Mathematical Approaches to Neural Networks
J.G. Taylor (Editor)

SELF–ORGANIZING NEURAL NETWORKS FOR STABLE CONTROL OF AUTONOMOUS BEHAVIOR IN A CHANGING WORLD

S. Grossberg‡

Department of Cognitive and Neural Systems, Boston University, Boston, MA, USA

1. INTRODUCTION: NONLINEAR MATHEMATICS FOR DESCRIBING AUTONOMOUS BEHAVIOR IN A NONSTATIONARY WORLD

The study of neural networks is challenging in part because the field embraces multiple goals. Neural networks to explain mind and brain are not evaluated by the same criteria as artificial neural networks for technology. Both are ultimately evaluated by their success in handling data, but data about behaving animals and humans may bear little resemblance to data that evaluates benchmark performance in technology. Although most artificial neural networks have been inspired by ideas gleaned from mind and brain models, technological applications can sometimes be carried out in an off-line setting with carefully selected data and complete external supervision. The living brain is, in contrast, designed to operate autonomously under real-time conditions in nonstationary environments that may contain unexpected events. Whatever supervision is available derives from the structure of the environment itself.

These facts about mind and brain subserve much of the excitement and the intellectual challenge of neural networks, particularly because many important applications need to be run autonomously in nonstationary environments that may contain unexpected events. What sorts of intuitive concepts are appropriate for analysing autonomous behavior that is capable of rapid adaptation to a changing world? What sorts of mathematics can express and analyse these concepts?

I have been fortunate to be one of the pioneers who has participated in the discovery and development of core concepts and models for the neural control of real–time autonomous behavior. A personal perspective on these developments will be taken in this chapter. Such a perspective has much to recommend it at this time. So many scientific communities and intellectual traditions have recently converged on the neural network field that a consistent historical viewpoint can simplify understanding.

When I began my scientific work as an undergraduate student in 1957, the modern field of neural networks did not exist. My main desire was to better understand how we

‡ This research was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225), DARPA (AFOSR 90-0083 and ONR N00014-92-J-4015), and the Office of Naval Research (ONR N00014-91-J-4100). The authors wish to thank Cynthia Bradford and Diana J. Meyers for their valuable assistance in the preparation of the manuscript.

humans manage to cope so well in a changing world. This required study of psychological data to become familiar with the visible characteristics of our behavioral endowment. It required study of neurobiological data to better understand how the brain is organized. New intuitive concepts and mathematical models were needed whereby to analyse these data and to link behavior to brain. New mathematical methods were sought to analyse how very large numbers of neural components interact over multiple spatial and temporal scales via nonlinear feedback interactions in real time. These methods needed to show how neural interactions may give rise to behaviors in the form of emergent properties.

Essentially no one at that time was trained to individually work towards all of these goals. Many experimentalists were superb at doing one type of psychological or neurobiological data, but rarely read broadly about other types of data. Few read across experimental disciplines. Even fewer knew any mathematics or models. The people who were starting to develop Artificial Intelligence favored symbolic mathematical methods. They typically disparaged the nonlinear differential equations that are needed to describe adaptive behavior in real time. Even the small number of people who used differential equations to describe brain or behavior often restricted their work to linear systems and avoided the use of nonlinear ones. It is hard to capture today the sense of overwhelming discouragement and ridicule that various of these people heaped on the discoveries of neural network pioneers. Insult was added to injury when their intellectual descendants eagerly claimed priority for these discoveries when they became fashionable years later. Their ability to do so was predicated on a disciplinary isolation of the psychological, neurobiological, mathematical, and computational communities that persisted for years after a small number of pioneers began their work to achieve an interdisciplinary synthesis.

Some of the historical factors that influenced the development of neural network research are summarized in Carpenter and Grossberg (1991) and Grossberg (1982a, 1987, 1988). The present discussion summarizes several contributions to understanding how neural models function autonomously in a stable fashion despite unexpected changes in their environments. The content of these models consists of a small set of equations that describe processes such as activation of short term memory (STM) traces, associative learning by adaptive weights or long term memory (LTM) traces, and slow habituative gating or medium term memory (MTM) by chemical modulators and transmitters; a larger set of modules that organize processes such as cooperation, competition, opponent processing, adaptive categorization, pattern learning, and trajectory formation; and a still larger set of neural systems or architectures for achieving general-purpose solutions of modal problems such as vision, speech, recognition learning, associative recall, reinforcement learning, adaptive timing, temporal planning, and adaptive sensory-motor control. Each successive level of model organization synthesizes several units from the previous level.

2. THE ADDITIVE AND SHUNTING MODELS

Two of the core neural network models that I introduced and mathematically analysed in their modern form are often called the additive model and the shunting model. These models were originally derived in 1957-1958 when I was an undergraduate at Dartmouth College. They describe how STM and LTM traces interact during network processes of activation, associative learning, and recall (Figure 1). It took ten years from their initial discovery and analysis to get them published in the intellectual climate of the 1960's

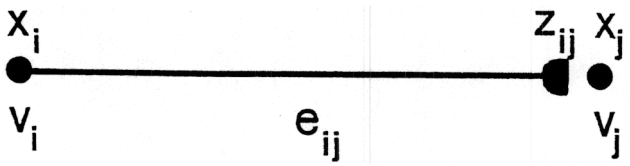


Figure 1. STM traces (or activities or potentials) x_i at cells (or cell populations) v_i emit signals along the directed pathways (or axons) e_{ij} which are gated by LTM memory traces (or adaptive weights) z_{ij} before they can perturb their target cells v_j. (Reprinted with permission from Grossberg, 1982c.)

(Grossberg, 1967, 1968a, 1968b). A monograph (Grossberg, 1964) that summarizes some of these results was earlier distributed to one hundred laboratories of leading researchers from the Rockefeller Institute where I was then a graduate student.

Additive STM Equation

$$\frac{d}{dt}x_i = -A_i x_i + \sum_{j=1}^{n} f_j(x_j)B_{ji}z_{ji}^{(+)} - \sum_{j=1}^{n} g_j(x_j)C_{ji}z_{ji}^{(-)} + I_i. \tag{1}$$

Equation (1) for the STM trace x_i includes a term for passive decay ($-A_i x_i$), positive feedback ($\sum_{j=1}^{n} f_j(x_j)B_{ji}z_{ji}^{(+)}$), negative feedback ($-\sum_{j=1}^{n} g_j(x_j)C_{ji}z_{ji}^{(-)}$), and input ($I_i$). Each feedback term includes a state-dependent nonlinear signal ($f_j(x_j), g_j(x_j)$), a connection, or path, strength (B_{ji}, C_{ji}), and an LTM trace ($z_{ji}^{(+)}, z_{ji}^{(-)}$). If the positive and negative feedback terms are lumped together and the connection strengths are lumped with the LTM traces, then the additive model may be written in the simpler form

$$\frac{d}{dt}x_i = -A_i x_i + \sum_{j=1}^{n} f_j(x_j)z_{ji} + I_i. \tag{2}$$

Early applications of the additive model included computational analyses in vision, associative pattern learning, pattern recognition, classical and instrumental conditioning, and the learning of temporal order in applications to language and sensory-motor control (Grossberg, 1969a, 1969b, 1969c, 1970a, 1970b, 1971a, 1972a, 1972b, 1974; Grossberg and Pepe, 1971). The additive model has continued to be a cornerstone of neural network research to the present day; see, for example, Amari and Arbib (1982) and Grossberg (1982a). Some physicists unfamiliar with the classical status of the additive model in neural network theory erroneously called it the Hopfield model after they became acquainted with Hopfield's first application of the additive model in Hopfield (1984), twenty-five years after its discovery; see Section 20. The classical McCulloch-Pitts (1943) model has also erroneously been called the Hopfield model by the physicists who became acquainted with the McCulloch-Pitts model in Hopfield (1982). These historical errors can ultimately be traced to the fact that many physicists and engineers who started studying neural networks in the 1980's generally did not know the field's scholarly literature. These errors are

gradually being corrected as new neural network practitioners learn the history of their craft.

A related network equation was found to more adequately model the shunting dynamics of individual neurons (Hodgkin, 1964; Kandel and Schwartz, 1981; Katz, 1966; Plonsey and Fleming, 1969). In such a shunting equation, each STM trace is restricted to a bounded interval $[-D_i, B_i]$ and automatic gain control, instantiated by multiplicative shunting terms, interacts with balanced positive and negative feedback signals and inputs to maintain the sensitivity of each STM trace within its interval.

Shunting STM Equation

$$\begin{aligned}\frac{d}{dt}x_i = &- A_i x_i + (B_i - x_i)[\sum_{j=1}^{n} f_j(x_j)C_{ji}z_{ji}^{(+)} + I_i] \\ &- (x_i + D_i)[\sum_{j=1}^{n} g_j(x_j)E_{ji}z_{ji}^{(-)} + J_i].\end{aligned} \tag{3}$$

Variations of the shunting equation (3) were also studied (Ellias and Grossberg, 1975) in which the reaction rate of inhibitory STM traces y_i was explicitly represented, as in the system

$$\begin{aligned}\frac{d}{dt}x_i = &- A_i x_i + (B_i - x_i)[\sum_{j=1}^{n} f_j(x_j)C_{ji}z_{ji}^{(+)} + I_i \\ &- (x_i + D_i)[\sum_{j=1}^{n} g_j(y_j)E_{ji}z_{ji}^{(-)} + J_i]\end{aligned} \tag{4}$$

and

$$\frac{d}{dt}y_i = -F_i y_i + (1 - G_i y_i)\sum_{j=1}^{n} f_j(x_j)H_{ij}. \tag{5}$$

Several LTM equations have been useful in applications. Two particularly useful variations have been:

Passive Decay LTM Equation

$$\frac{d}{dt}z_{ij} = -K_{ij}z_{ij} + L_{ij}f_i(x_i)h_j(x_j) \tag{6}$$

and

Gated Decay LTM Equation

$$\frac{d}{dt}z_{ij} = h_j(x_j)[-K_{ij}z_{ij} + L_{ij}f_i(x_i)]. \tag{7}$$

In both equations, a nonlinear learning term $f_i(x_i)h_j(x_j)$, often called a Hebbian term after Hebb (1949), is balanced by a memory decay term. In (6), memory decays passively at a constant rate $-K_{ij}$. In (7), memory decay is gated on and off by one of the nonlinear signals. When the gate opens, z_{ij} tracks $f_i(x_i)$ by steepest descent. A key property of

both equations is that the size of an LTM trace z_{ij} can either increase or decrease due to learning.

Neurophysiological support for an LTM equation of the form (7) was reported two decades after it was first introduced (Levy, 1985; Levy, Brassel, and Moore, 1983; Levy and Desmond, 1985; Rauschecker and Singer, 1979; Singer, 1983). Extensive mathematical analyses of these STM and LTM equations in a number of specialized circuits led gradually to the identification of a general class of networks for which one could prove invariant properties of associative spatiotemporal pattern learning and recognition (Grossberg, 1969a, 1971b, 1972c, 1982). These mathematical analyses helped to identify those features of the models that led to useful emergent properties. They sharpened intuition by showing the implications of each idea when it was realized within a complex system of interacting components. Some of these results are summarized below.

3. UNITIZED NODES, SHORT TERM MEMORY, AND AUTOMATIC ACTIVATION

The neural network framework and the additive laws were derived in several ways (Grossberg, 1969a, 1969b, 1969f, 1974). My first derivation in 1957–1958 was based on classical list learning data (Grossberg, 1961, 1964) from the serial verbal learning and paired associate paradigms (Dixon and Horton, 1968; Jung, 1968; McGeogh and Irion, 1952; Osgood, 1953; Underwood, 1966). List learning data force one to confront the fact that new verbal units are continually being synthesized as a result of practice, and need not be the obvious units which the experimentalist is directly manipulating (Young, 1968). All essentially stationary concepts, such as the concept of information itself (Khinchin, 1967) hereby became theoretically useless.

By putting the self-organization of individual behavior in center stage, I realized that the phenomenal simplicity of familiar behavioral units, and the synthesis of these units into new representations which themselves achieve phenomenal simplicity through experience, should be made a fundamental property of the theory. To express the phenomenal simplicity of familiar behavioral units, I represented them by indecomposable internal representations, or unitized nodes, $v_i, i = 1, 2, \ldots, n$. This hypothesis gained support from the (now classical) paper of Miller (1956) on the Magic Number Seven, which appeared at around the time I was doing this derivation. In this work, Miller described how composites of familiar units can be "chunked", or unitized, into new units via the learning process. Miller used the concept of information to analyse his results. This concept cannot, however, be used to explain how chunking occurs. A neural explanation of the Magic Number Seven is described in Grossberg (1978a, 1986); see also Cohen and Grossberg (1986).

Data concerning the manner in which humans learn serial lists of verbal items led to the first derivation of the additive model. These data were particularly helpful because the different error distributions and learning rates at each list position suggested how each list item dynamically senses and learns from a different spatiotemporal context. It was, for example, known that practicing a list of items such as AB could also lead to learning of BA, a phenomenon called backward learning. A list such as ABC can obviously also be learned, however, showing that the context around item B enables forward learning of BC to supercede backward learning of BA.

To simplify the discussion of such interactive phenomena, I will consider only associative interactions within a given level in a coding hierarchy, rather than the problem of how coding hierarchies develop and interact between several levels. All of these conclusions have been generalized to a hierarchical setting (Grossberg, 1974, 1978a, 1980a).

4. BACKWARD LEARNING AND SERIAL BOWING

Backward learning effects and, more generally, error gradients between nonadjacent, or remote, list items (Jung, 1968; McGeogh and Irion, 1952; Murdock, 1974; Osgood, 1953; Underwood, 1966) suggested that pairs of nodes v_i and v_j can interact via distinct directed pathways e_{ij} and e_{ji} over which adaptive signals can travel. An analysis of how a node v_i could know where to send its signals revealed that no local information exists at the node itself whereby such a decision could be made. By the principle of sufficient reason, the node must therefore send signals towards all possible nodes v_j with which it is connected by directed paths e_{ij}. Some other variable must exist that discriminates which combination of signals can reach their target nodes based on past experience. These auxiliary variables turned out to be the long term memory traces. The concept that each node sends out signals to all possible nodes subsequently appeared in models of *spreading activation* (Collins and Loftus, 1975; Klatsky, 1980) to explain semantic recognition and reaction time data.

The form that the signaling and learning laws should take was suggested by data about serial verbal learning. During serial learning, a subject is presented with one list item at a time and asked to predict the next item before it occurs. After a rest period, the list is presented again. This procedure continues until a fixed learning criterion is reached. A main paradox about serial learning concerns the form of the bowed serial position curve which relates cumulative errors to list positions (Figure 2a). This curve is paradoxical for the following reason. If all that happened during serial learning was a build-up of interference at each list position due to the occurrence of prior list items, then the error curve should be monotone increasing (Figure 2b). Because the error curve is bowed, and the degree of bowing depends on the length of the rest period, or intertrial interval, between successive list presentations, the *nonoccurrence* of list items after the last item occurs somehow improves learning across several prior list items. Internal events thus continue to occur during the intertrial interval. The nonoccurrence of future items can hereby reorganize the learning of a previously occurring list.

The bowed serial position curve showed me that a real-time dynamical theory was needed to understand how these internal events continue to occur even after external inputs cease. It also showed that these internal events can somehow operate "backwards in time" relative to the external ordering of observable list items. These backward effects suggested that directed network interactions exist whereby a node v_i could influence a node v_j, and conversely.

Many investigators attributed properties like bowing to one or another kind of rehearsal (Klatsky, 1980; Rundus, 1971). Just saying that rehearsal causes bowing does not explain it, because it does not explain why the middle of the list is less rehearsed. Indeed the middle of the list has more time to be rehearsed than does the end of the list before the next learning trial occurs. In the classical literature, the middle of the list was also said to experience maximal proactive interference (from prior items) and retroactive interference (from future items), but this just labels what we have to explain

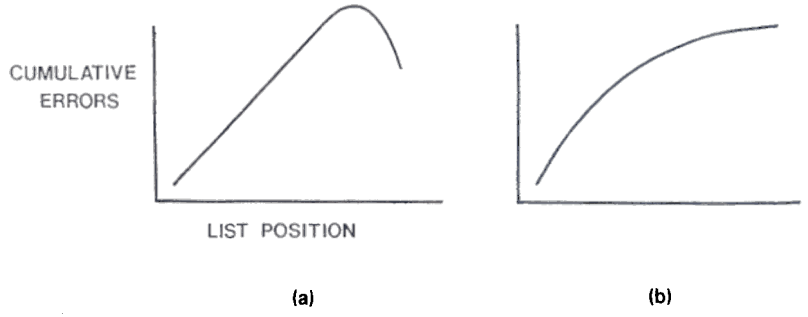


Figure 2. (a) The cumulative error curve in serial verbal learning is a skewed bowed curve. Items between the middle and end of the list are hardest to learn. Items at the beginning of the list are easiest to learn. (b) If position-dependent difficulty of learning were all due to interference from previously presented items, the error curve would be monotone increasing. (Reprinted with permission from Grossberg, 1982b.)

(Osgood, 1953; Underwood, 1966). The severity of such difficulties led the serial learning expert Young (1968) to write: "If an investigator is interested in studying verbal learning processes ... he would do well to choose some method other than serial learning" (p.146). Another leading verbal learning expert Underwood (1966) wrote: "The person who originates a theory that works out to almost everyone's satisfaction will be in line for an award in psychology equivalent to the Nobel prize" (p. 491). It is indicative of the isolated role of real-time modelling in psychology at that time that a theory capable of clarifying the main data effects was available but could not yet get published. Similar chunking and backward effects also occur in a wide variety of problems in speech, language, and adaptive sensory-motor control, so avoiding serial learning will not make the problem go away. Indeed these phenomena may all generally be analysed using the same types of mechanisms.

5. THE NEED FOR A REAL-TIME NETWORK THEORY

The massive backward effect that causes the bowed serial curve forced the use of a real-time theory that can parameterize the temporal unfolding of both the occurrences and the nonoccurrences of events. The existence of facilitative effects due to nonoccurring items also showed that traces of prior list occurrences must endure beyond the last item's presentation time, so they can be influenced by the future nonoccurrences of items. This fact led to the concept of activations, or short term memory (STM) traces, $x_i(t)$ at the nodes v_i, $i = 1, 2, \ldots, n$, which are turned on by inputs $I_i(t)$, but which decay at a rate slower than the input presentation rate. As a result, in response to serial inputs, *patterns* of STM activity are set up across the network's nodes. The combination of serial inputs, distributed internodal signals, and spontaneous STM changes at each node changes the STM pattern as the experiment proceeds. A major task of neural network theory was thus to learn how to think in terms of distributed pattern transformations, rather than just in terms of distributed feature detectors or other local entities. When I first realized this, it was quite a radical notion. Now it is so taken for granted that most people do not

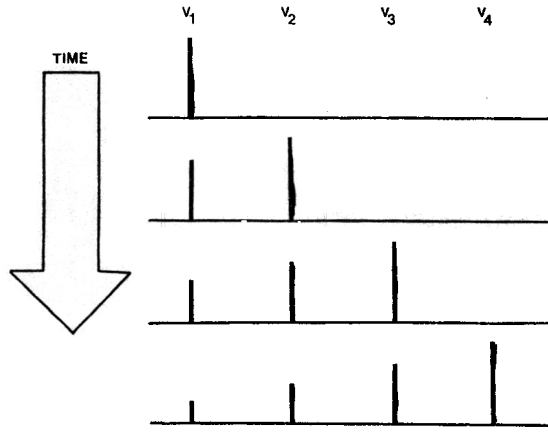


Figure 3. Suppose that items $r_1, r_2, r_3, r_4, \ldots$ are presented serially to nodes v_1, v_2, v_3, $v_4, \ldots$, respectively. Let the activity of node v_i at time t be described by the height of the histogram beneath v_i at time t. If each node is initially excited by an equal amount and its excitation decays at a fixed rate, then at every time (each row) the pattern of STM activity across nodes is described by a recency gradient. (Reprinted with permission from Grossberg, 1982b.)

realize that is was once an exciting discovery.

6. EVENT TEMPORAL ORDER VS. LEARNED TEMPORAL ORDER

The general philosophical interest of the bowed error curve can be appreciated by asking: What is the first time a learning subject can possibly know that item r_n is the last list item in a newly presented list $r_1 r_2 \ldots r_n$, given that a new item is presented every w time units until r_n occurs? The answer obviously is: not until at least w time units *after* r_n has been presented. Only after this time passes and no item r_{n+1} is presented can r_n be correctly reclassified from the list's "middle" to the list's "end". The nonoccurrence of future items reclassifies r_n as the "end" of the list. Parameter w is under experimental control and is not a property of the list ordering *per se*. Spatiotemporal network interactions thus parse a list in a way that is fundamentally different from the parsing rules that are natural to apply to a list of symbols in a computer. Indeed, increasing the event presentation rate, or intratrial interval, w during serial learning can flatten the entire bowed error curve and minimize the effects of the intertrial interval between successive list presentations (Jung, 1968; Osgood, 1953).

To illustrate further the difference between computer models and a real-time network approach, suppose that after a node v_i is excited by an input I_i, its STM trace gets smaller through time due to either internodal competition or to passive trace decay. Then in response to a serially presented list, the last item to occur always has the largest STM trace—in other words, at every time a *recency* gradient obtains in STM (Figure 3). Given this natural assumption—which, however, is not always true (Bradski, Carpenter, and Grossberg, 1992; Grossberg, 1978a, 1978b)—how do the generalization gradients of

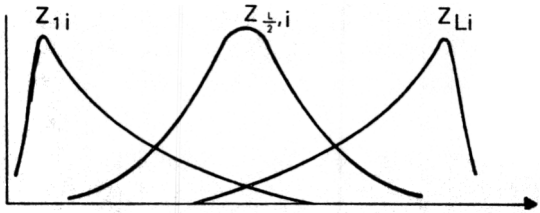


Figure 4. At each node v_j, the LTM pattern $z_j = (z_{j1}, z_{j2}, \ldots, z_{jn})$ that evolves through time is different. In a list of length $n = L$ whose intertrial interval is sufficiently long, the LTM pattern at the list beginning ($j \cong 1$) is a primacy gradient. At the list end ($j \cong L$), a recency gradient evolves. Near the list middle ($j \cong \frac{L}{2}$), a two-sided gradient is learned. These gradients are reflected in the distribution of anticipatory and perseverative errors in response to item probes at different list positions. (Reprinted with permission from Grossberg, 1982b.)

errors at each list position get learned (Figure 4)? In particular, how does a gradient of anticipatory, or forward, errors occur at the beginning of the list, a gradient of perseverative, or backward, errors occur at the end of the list and a two-sided gradient of anticipatory and perseverative errors occur near the middle of the list (Osgood, 1953)? Otherwise expressed, how does a temporal succession of STM recency gradients generate an LTM *primacy* gradient at the list beginning but an LTM *recency* gradient at the list end? I call this *STM–LTM order reversal.* This property immediately rules out any linear theory, as well as any theory which restricts itself to nearest neighbor associative links.

7. MULTIPLICATIVE SAMPLING BY SLOWLY DECAYING LTM TRACES OF RAPIDLY EVOLVING STM PATTERNS

The STM and LTM properties depicted in Figures 3 and 4 can be reconciled by positing the existence of STM traces and LTM traces that evolve according to different time scales and rules. Indeed, this reconciliation was one of the strongest arguments that I knew for these rules until neurobiological data started to support them during the 1980's.

Suppose that the STM trace of each active node v_j can send out a sampling signal S_j along each directed path e_{jk} towards the node v_k, $k \neq j$. Suppose that each path e_{jk} contains LTM trace z_{jk} at its terminal point, where z_{jk} can compute, using only local operations, the product of signal S_j and STM trace x_k. Also suppose that the LTM trace decays slowly, if at all, during a single learning trial. The simplest law for z_{jk} that satisfies these constraints is

$$\frac{d}{dt} z_{jk} = -cz_{jk} + dS_j x_k, \tag{8}$$

$j \neq k$; cf., equation (6). To see how this rule generates an LTM primacy gradient at the list beginning, we need to study the LTM pattern $(z_{12}, z_{13}, \ldots, z_{1n})$ and to show that $z_{12} > z_{13} > \ldots > z_{1n}$. To see how the same rule generates an LTM recency gradient at the list end, we need to study the LTM pattern $(z_{n1}, z_{n2}, \ldots, z_{n,n-1})$ and to show that $z_{n1} < z_{n2} < \ldots < z_{n,n-1}$. The two-sided gradient at the list middle can then be understood as a combination of these effects.

By (8), node v_1 sends out a sampling signal S_1 shortly after item r_1 is presented. After rapidly reaching peak size, signal S_1 gradually decays as future list items $r_2, r_3, \ldots$ are presented. Thus S_1 is largest when trace x_2 is maximal, S_1 is smaller when both traces x_2 and x_3 are active, S_1 is smaller still when traces x_2, x_3, and x_4 are active, and so on. Consequently, the product S_1x_2 in row 2 of Figure 3 exceeds the product S_1x_3 in row 3 of Figure 3, which in turn exceeds the product S_1x_4 in row 4 of Figure 3, and so on. Due to the slow decay of each LTM trace z_{1k} on each learning trial, z_{12} adds up to the products S_1x_2 in successive rows of column 1, z_{13} adds up to the products S_1x_3 in successive rows of column 2, and so on. An LTM primacy gradient $z_{12} > z_{13} > \ldots > z_{1n}$ is hereby generated. This gradient is due to the way signal S_1 *multiplicatively samples* the successive STM recency gradients and the LTM traces z_{1k} sum up the sampled STM gradients.

By contrast, the signal S_n of a node v_n at the end of the list samples a different set of STM gradients. This is because v_n starts to sample (viz., $S_n > 0$) only after all past nodes $v_1, v_2, \ldots, v_{n-1}$ have already been activated on that trial. Consequently, the LTM traces $(z_{n1}, z_{n2}, \ldots, z_{n,n-1})$ of node v_n encode a recency gradient $x_1 < x_2 < x_3 < \ldots < x_{n-1}$ at *each* time. When all the recency gradients are added up through time, the total effect is a recency gradient in v_n's LTM pattern. In summary, nodes at the beginning, middle, and end of the list encode different LTM gradients because they multiplicatively sample and store STM patterns at different times. Similar LTM gradients obtain if the sequences of nodes which are active at any time selectively excite higher-order nodes, or chunks, which in turn sample the field of excited nodes via feedback signals (Grossberg, 1974, 1978a).

8. MULTIPLICATIVE LTM GATING OF STM-ACTIVATED SIGNALS

Having shown how STM patterns may be read into LTM patterns, we now need to describe how a retrieval probe r_m can read v_m's LTM pattern back into STM on recall trials, whereupon some of the STM traces can be transformed into observable behavior. In particular, how can LTM be read into STM without distorting the learned LTM gradients?

The simplest rule generates an STM pattern which is proportional to the LTM pattern that is being read out, and allows distinct probes to each read their LTM patterns into STM in an independent fashion. To achieve faithful read-out of the LTM pattern $(z_{m1}, z_{m2}, \ldots, z_{mn})$ by a probe r_m that turns on signal S_m, let the product $S_m z_{mi}$ determine the growth rate of x_i. Then LTM trace z_{mi} *gates* the signal S_m along e_{mi} before the gated signal reaches v_i. The independent action of several probes implies that the gated signals $S_m z_{mi}$ are added, so that the total effect of all gated signals on v_i is $\sum_{m=1}^{n} S_m z_{mi}$. The simplest equation for the STM trace x_i that abides by this rule is the additive equation

$$\frac{d}{dt}x_i = -ax_i + b\sum_{m=1}^{n} S_m z_{mi} + I_i, \tag{9}$$

where $-a$ is the STM decay rate, S_m is the mth sampling signal, z_{mi} is the LTM trace of pathway e_{mi}, and I_i is the ith experimental input; cf, equation (2).

The reaction of equations (8) and (9) to serial inputs I_i is much more complex than is their response to an isolated retrieval probe r_m. Due to the fact that STM traces may decay slower than the input presentation rate, several sampling signals S_m can be simultaneously active, albeit in different phases of their growth and decay. In fact, this in-

teraction leads to properties that mimmick list learning data, but first a technical problem needs to be overcome.

9. BEHAVIORAL CHOICE AND COMPETITIVE INTERACTIONS

Once one accepts that patterns of STM traces are evolving through time, one also needs a mechanism for choosing those activated nodes which will influence observable behavior. Lateral inhibitory feedback signals were derived as a choice mechanism (Grossberg, 1968, 1969b, 1970a). The simplest extension of (9) which includes competitive interactions is

$$\frac{d}{dt}x_i = -ax_i + \sum_{m=1}^{n} S_m^+ b_{mi}^+ z_{mi} - \sum_{m=1}^{n} S_m^- b_{mi}^- + I_i \tag{10}$$

where $S_m^+ b_{mi}^+$ ($S_m^- b_{mi}^-$) is the excitatory (inhibitory) signal emitted from node v_m along the excitatory (inhibitory) pathway e_{mi}^+ (e_{mi}^-); cf., equation (1). Correspondingly equation (8) is generalized to

$$\frac{d}{dt}z_{jk} = -cz_{jk} + d_{jk}S_j^+ x_k. \tag{11}$$

The asymmetry between terms $\sum_{m=1}^n S_m^+ b_{mi}^+ z_{mi}$ and $\sum_{m=1}^n S_m^- b_{mi}^-$ in (10) suggested a modification of (10) and a definition of inhibitory LTM traces analogous to the excitatory LTM traces (8), where such inhibitory traces exist (Grossberg, 1969d).

Because lateral inhibition can change the sign of each x_i from positive to negative in (10), and thus change the sign of each z_{jk} from positive to negative in (8), some refinements of (10) and (8) were needed to prevent absurdities like the following: $S_m^+ < 0$ and $x_i < 0$ implies $z_{mi} > 0$; and $S_m^+ < 0$ and $z_{mi} < 0$ implies $x_i > 0$. Signal thresholds accomplished this in the simplest way. Letting $[\xi]^+ = \max(\xi, 0)$, define the threshold-linear signals.

$$S_j^+ = [x_j(t - \tau_j^+) - \Gamma_j^+]^+ \tag{12}$$

and

$$S_j^- = [x_j(t - \tau_j^-) - \Gamma_j^-]^+, \tag{13}$$

in (10) and (11), and modify (10) to read

$$\frac{d}{dt}z_{jk} = -cz_{jk} + d_{jk}S_j^+[x_k]^+. \tag{14}$$

Sigmoid, or S-shaped signals, were also soon mathematically shown to support useful computational properties (Grossberg, 1973). These additive equations and their variants have been used by many subsequent modellers.

10. THE SKEWED BOW: SYMMETRY-BREAKING BETWEEN FUTURE AND PAST

One of the most important contributions of neural network models has been to show how behavioral properties can arise as emergent properties due to network interactions. The bowed error curve is perhaps the first behaviorally important emergent property that was derived from a real-time neural network. It results from forward and backward interactions among all the STM and LTM variables across the network.

To explain the bowed error curve, we need to compare the LTM patterns $z_i = (z_{i1}, z_{i2}, \ldots, z_{in})$ that evolve at all list nodes v_i. In particular, we need to explain why the bowed curve is *skewed*; that is, why the list position where learning takes longest occurs nearer to the end of the list than to its beginning (Figure 2a). This skewing effect contradicts learning theories that assume forward and backward effects are equally strong, or symmetric (Asch and Ebenholtz, 1962; Murdock, 1974). This symmetry-breaking between the future and the past, by favoring forward over backward associations, makes possible the emergence of a global "arrow in time," or the ultimate learning of long event sequences in their correct order, much as we learn the alphabet ABC ... Z despite the existence of backward learning.

A skewed bowed error curve does emerge in the network, and predicts that the degree of skewing will decrease, and the relative learning rate at the beginning and end of the list will reverse, as the network's arousal level increases or its signal thresholds Γ_j^+ decrease to abnormal levels (Grossberg and Pepe, 1971). The arousal and threshold predictions have not yet been directly tested to the best of my knowledge. Abnormally high arousal or low thresholds generate a formal network syndrome characterized by contextual collapse, reduced attention span, and fuzzy response categories that resemble aspects of simple schizophrenia (Grossberg and Pepe, 1970; Maher, 1977).

To understand intuitively what is involved in this explanation of bowing, note that by equation (14), each correct LTM trace $z_{12}, z_{23}, z_{34}, \ldots, z_{n-1,n}$ that is activated by list item r_1 may grow at a comparable rate, albeit w time units later than the previous correct LTM trace. However, the LTM patterns $z_1, z_2, \ldots, z_n$ differ at every list position, as in Figure 4. Thus when a retrieval probe r_j reads its LTM pattern z_j into STM, the entire pattern must influence overt behavior to explain why bowing occurs. The *relative* size of the correct LTM trace $z_{j,j+1}$ compared to all other LTM traces in z_j will influence its success in eliciting r_{j+1} after competitive STM interactions occur. A larger $z_{j,j+1}$ relative to the sum of all other z_{jk}, $k \neq j, j+1$, should yield better performance of r_{j+1} given r_j, other things being equal. To measure the distinctiveness of a trace z_{jk} relative to all traces in z_j, I therefore defined the relative LTM traces

$$Z_{jk} = z_{jk}(\sum_{m \neq j} z_{jm})^{-1}. \tag{15}$$

Equation (15) provides a convenient measure of the effect of LTM on STM after competition acts. By (15), the ordering within the LTM gradients of Figure 4 is preserved by the relative LTM traces; for example, if $z_{12} > z_{13} > \ldots > z_{1n}$, then $Z_{12} > Z_{13} > \ldots > Z_{1n}$ because all the Z_{1k}'s have the same denominator. Thus all conclusions about LTM gradients are valid for relative LTM gradients, which are also sometimes called stimulus sampling probabilities.

In terms of the relative LTM traces, the issue of bowing can be mathematically formulated as follows. Define the *bowing function* $B_i(t) = Z_{i,i+1}(t)$. Function $B_i(t)$ measures how distinctive the ith correct association is at time t. After a list of n items is presented with an intratrial interval w and a sufficiently long intertrial interval W elapses, does the function $B_i((n-1)w + W)$ decrease and then increase as i increases from 1 to n? Does the minimum of the function occur in the latter half of the list? The answer to both of these questions is "yes."

To understand why this happens, it is necessary to understand how the bow depends upon the ability of a node v_i to sample incorrect future associations, such as $r_i r_{i+2}, r_i r_{i+3}, \ldots$ in addition to incorrect past associations, such as $r_i r_{i-1}, r_i r_{i-2}, \ldots$. As soon as S_i becomes positive, v_i can sample the entire past field of STM traces at $v_1, v_2, \ldots, v_{i-1}$. However, if the sampling threshold is chosen high enough, S_i might shut off before r_{i+2} occurs. Thus the sampling duration has different effects on the sampling of past than of future incorrect associations. For example, if the sampling thresholds of all v_i are chosen so high that S_i shuts off before r_{i+2} is presented, then the function $B_i(\infty)$ decreases as i increases from 1 to n. In other words, the monotonic error curve of Figure 2b obtains because no node v_i can encode incorrect future associations. Even if the thresholds are chosen so that incorrect future associations can be formed, the function $B_i((i+1)w)$ which measures the distinctiveness of $z_{i,i+1}$ just before r_{i+2} occurs is again a decreasing function of i. The bowing effect thus depends on threshold choices which permit sampling durations that are at least $2w$ in length.

The shape of the bow also depends on the duration of the intertrial interval, because before the intertrial interval occurs, all nodes build up increasing amounts of associative interference as more list items are presented. The first effect of the nonoccurrence of items after r_n is presented is the growth through time of $B_{n-1}(t)$ as t increases beyond the time nw when item r_{n+1} would have occurred in a larger list. The last correct association is hereby facilitated by the absence of interfering future items during the intertrial interval. This facilitation effect is a nonlinear property of the network. Bowing is also a nonlinear phenomenon in the theory, because it depends on a comparison of ratios of integrals of sums of products as they evolve through time.

Mathematical theorems about the bowed error curve and other list learning properties were described in Grossberg (1969c) and Grossberg and Pepe (1971), and reviewed in Grossberg (1982a, 1982b). These results illustrated how STM and LTM processes interact as unitized events occur sequentially in time. Other mathematical studies analysed increasingly general constraints under which distributed STM patterns could be encoded in LTM without bias by arbitrary numbers of simultaneously active sampling nodes acting in parallel. Some of these results are summarized in the next section.

11. ABSOLUTELY STABLE PARALLEL PATTERN LEARNING

Many features of system (10) and (12)–(14) are special; for example, the exponential decay of STM and LTM and the signal threshold rule. Because associative processing is ubiquitous throughout phylogeny and within functionally distinct subsystems of each individual, a more general mathematical framework was needed. This framework needed to distinguish universally occurring associative principles that guarantee essential learning properties from evolutionary variations that adapt these principles to realize specialized skills.

I approached this problem from 1967 to 1972 in a series of articles wherein I gradually realized that the mathematical properties used to globally analyze specific learning examples were much more general than the examples themselves. This work culminated in my universal theorems on associative learning (Grossberg, 1969d, 1971a, 1972a). The theorems say that if certain associative laws were invented at a prescribed time during evolution, then they could achieve unbiased associative pattern learning in essentially any

later evolutionary specialization. To the question: Was it necessary to re-invent a new learning rule to match every perceptual or cognitive refinement, the theorems said "no". They enabled arbitrary spatial patterns to be learned by arbitrarily many, simultaneously active sampling channels that are activated by arbitrary continuous data preprocessing in an essentially arbitrary anatomy. Arbitrary space-time patterns can also be learned given modest constraints on the temporal regularity of stimulus sampling. The universal theorems thus describe a type of parallel processing whereby unbiased associative pattern learning occurs despite mutual crosstalk between nonlinear feedback signals.

These results obtain only if the network's main computations, such as spatial averaging, temporal averaging, preprocessing, gating, and cross-correlation are computed in a canonical ordering. This canonical ordering constitutes a general purpose design for unbiased parallel pattern learning, as well as a criterion for whether particular networks are acceptable models for this task. The universality of the design mathematically takes the form of a classification of oscillatory and limiting possibilities that is invariant under evolutionary specializations.

The theorems can also be interpreted in another way that is appropriate in discussions of self-organizing systems. The theorems are *absolute stability* or *global content addressable memory* theorems. They show that evolutionary invariants of associative learning obtain no matter how system parameters are changed within this class of systems. Absolutely stable learning is an important property in a self-organizing system because parameters may change in ways that cannot be predicted in advance, notably when unexpected environments act on the system. Absolute stability guarantees that the onset of self-organization does not subvert the very learning properties that make stable self-organization possible.

The systems that I considered constitute the *generalized additive model*

$$\frac{d}{dt}x_i = -A_i x_i + \sum_j B_{ji} z_{ji} + I_i(t) \tag{16}$$

$$\frac{d}{dt}z_{ji} = -C_{ji} z_{ji} + D_{ji} x_i \tag{17}$$

where i and j parameterize arbitrarily large, not necessarily disjoint, sets of sampled and sampling cells, respectively. As in my equations for list learning, A_i is an STM decay rate, B_{ki} is a nonnegative performance signal, $I_i(t)$ is an input function, C_{ji} is an LTM decay rate, and D_{ji} is a nonnegative learning signal. Unlike the list learning equations, A_i, B_{ki}, C_{ji}, and D_{ji} may be continuous functionals of the entire history of the system. Equations (16) and (17) are thus very general, and include many of the specialized associative learning models in the literature.

For example, although (16) does not seem to include inhibitory interactions, such interactions may be lumped (say) into the STM decay functional A_i. The choice

$$A_i = a_i - (b_i - c_i x_i) G_i(x_i) + \sum_{k=1}^{n} H_k(x_k) d_{ki} \tag{18}$$

describes the case wherein system nodes compete via shunting, or membrane equation, interactions (Cole, 1968; Grossberg, 1973; Kandel and Schwartz, 1981; Plonsey and Fleming,

1969). The performance, LTM decay, and learning functionals may include slow threshold changes, nonspecific Now Print signals, signal velocity changes, presynaptic modulation, arbitrary continuous rules of dendritic preprocessing and axonal signaling, as well as many other possibilities (Grossberg, 1972a, 1974). Of special importance are the variety of LTM decay choices that satisfy the theorems. For example, a gated LTM law like

$$\frac{d}{dt}z_{ji} = [x_j(t-\tau_j) - \Gamma_j(y_t)]^+(-d_j z_{ji} + e_j x_i) \tag{19}$$

achieves an interference theory of forgetting, rather than exponential forgetting, since $\frac{d}{dt}z_{ji} = 0$ except when v_j is sampling (Adams, 1967); cf., equation (7). Equation (19) also allows the vigor of sampling to depend on changes in the threshold $\Gamma_j(y_t)$ that are sensitive to the prior history $y_t = (x_i, z_{ji} : i \in I, j \in J)_t$ of the system before time t, as in the model of Bienenstock, Cooper, and Munro (1982).

In this generality, too many possibilities exist to as yet prove absolute stability theorems. Indeed, if the performance signals B_{ji} from a fixed sampling node v_j to all the sampled nodes v_i, $i \in I$, were arbitrary nonnegative and continuous functionals, then the irregularities in each B_{ji} could override any regularities in z_{ji} within the gated performance signal $B_{ji}z_{ji}$ from v_j to v_i. One further constraint was used to impose some spatiotemporal regularity on the sampling process, as indicated in the next section.

12. LOCAL SYMMETRY, ACTION POTENTIALS, AND UNBIASED LEARNING

Absolute stability obtains even if different functionals B_j, C_j, and D_j are assigned to each node v_j, $j \in J$, just so long as the same functional is assigned to all pathways e_{ji}, $i \in I$. Where this is not globally true, one can often partition the network into maximal subsets where it is true, and then prove unbiased pattern learning in each subset. This restriction is called the property of *local symmetry axes* since each sampling cell v_j can act as a source of coherent history-dependent waves of STM and LTM processing. Local symmetry axes still permit (say) each B_j to obey different history-dependent preprocessing, threshold, time lag, and path strength laws among arbitrarily many mutually interacting nodes v_j.

When local symmetry axes are imposed on the generalized additive model in (16) and (17), the resulting class of systems takes the form

$$\frac{d}{dt}x_i = -Ax_i + \sum_{k\in J} B_k z_{ki} + I_i(t) \tag{20}$$

and

$$\frac{d}{dt}z_{ji} = -C_j z_{ji} + D_j x_i. \tag{21}$$

A change of variables shows, moreover, that constant interaction coefficients b_{ji} between pairs v_j and v_i of nodes can depend on $i \in I$ without destroying unbiased pattern learning in the systems

$$\frac{d}{dt}x_i = -Ax_i + \sum_j B_j b_{ji} z_{ji} + I_i(t) \tag{22}$$

and

$$\frac{d}{dt}z_{ji} = -C_j z_{ji} + D_j b_{ji}^{-1} x_i. \tag{23}$$

By contrast, the systems (22) and

$$\frac{d}{dt}z_{ji} = -C_j z_{ji} + D_j b_{ji} x_i \tag{24}$$

are not capable of unbiased parallel pattern learning (Grossberg, 1972a). A dimensional analysis showed that (22) and (23) hold if action potentials transmit the network's intercellular signals, whereas (22) and (24) hold if electrotonic propagation is used. The cellular property of an action potential was hereby formally linked to the network property of unbiased parallel pattern learning.

13. THE UNIT OF LTM IS A SPATIAL PATTERN

These global theorems proved that "the unit of LTM is a spatial pattern". This result was surprising to me, even though I had discovered the additive model. The result illustrates how rigorous mathematics can force insights that go beyond unaided intuition. In the present instance, it suggested a new definition of spatial pattern and showed how the network learns "temporally coherent spatial patterns" that may be hidden in its distributed STM activations through time. This theme of temporal coherence, first mathematically discovered in 1966, has shown itself in many forms since, particularly in recent studies of attention, resonance, and synchronous oscillations (Crick and Koch, 1990; Eckhorn, Bauer, Jordan, Brosch, Kruse, Munk, and Reitbock, 1988; Eckhorn and Schanze, 1991; Gray and Singer, 1989; Gray, Konig, Engel, and Singer, 1989; Grossberg, 1976c; Grossberg and Somers, 1991, 1992).

To illustrate the global theorems that have been proved, I consider first the simplest case, wherein only one sampling node v_0 exists (Figure 5a). Then the network is called an *outstar* because it can be drawn with the sampling node at the center of outward-facing adaptive pathways (Figure 5b) such that the LTM trace z_{0i} in the ith pathway samples the STM trace x_i of the ith sampled cell, $i \in I$. An *outstar* is thus a neural network of the form

$$\frac{d}{dt}x_i = -Ax_i + Bz_{0i} + I_i(t) \tag{25}$$

$$\frac{d}{dt}z_{0i} = -Cz_{0i} + Dx_i \tag{26}$$

where A, B, C, and D are continuous functionals such that B and E are nonnegative.

Despite the fact that the functionals A, B, C, and D can fluctuate in complex system-dependent ways, and the inputs $I_i(t)$ can also fluctuate wildly through time, an outstar can learn an arbitrary *spatial pattern*

$$I_i(t) = \theta_i I(t)$$

where $\theta_i \geq 0$ and $\sum_{k \in I} \theta_k = 1$, with a minimum of oscillations in its pattern variables $X_i = x_i(\sum_{k \in I} x_k)^{-1}$ and $Z_i = z_i(\sum_{k \in I} z_k)^{-1}$. These pattern variables learn the temporally coherent weights θ_i in a spatial pattern and factor the input activation $I(t)$ that energizes the process into the learning rate. The Z_i's are the relative LTM traces (15) that played such a central role in the explanation of serial bowing. The limiting and oscillatory behaviors of the pattern variables have a classification that is independent of particular

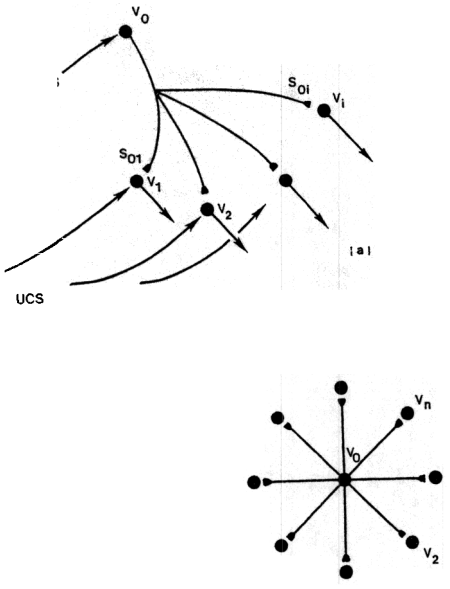


Figure 5. (a) The minimal anatomy capable of associative learning. For example, during classical conditioning, a conditioned stimulus (CS) excites a single node, or cell population, v_0 which thereupon sends sampling signals to a set of nodes $v_1, v_2, \ldots, v_n$. An input pattern representing an unconditioned stimulus (UCS) excites the nodes $v_1, v_2, \ldots, v_n$, which thereupon elicit output signals that contribute to the unconditioned response (UCR). The sampling signals from v_0 activate the LTM traces z_{0i} $i = 1, 2, \ldots, n$. The activated LTM traces can learn the activity pattern across $v_1, v_2, \ldots, v_n$ that represents the UCS. (b) When the sampling structure in (a) is redrawn to emphasize its symmetry, the result is an *outstar*, whose sampling source is v_0 and whose sampled border is the set of nodes $\{v_1, v_2, \ldots, v_n\}$. (Reprinted with permission from Grossberg, 1982b.)

choices of A, B, C, D, and I. These properties are thus evolutionary invariants of outstar learning.

The following theorem summarizes, albeit not in the most general known form, some properties of outstar learning. One of the constraints in this theorem is called a *local flow* condition. This constraint says that a performance signal B can be large only if its associated learning signal D is large. When local flow holds, pathways which have lost their plasticity can be grouped into the total input pattern that is registered in STM for encoding in LTM by other pathways.

If the threshold of the performance signal B is no smaller than the threshold of the learning signal D, then local flow is assured. Such a threshold inequality occurs automatically if the LTM trace z_{ji} is physically interpolated between the axonal signal and the postsynaptic target cell v_i. That is why the condition is called a local flow condition.

Such a geometric interpretation of the location of the LTM trace gives unexpected support to the hypothesis that LTM traces are localized in the synaptic knobs or postsynaptic membranes of cells undergoing associative learning. Here again a network property gives new functional meaning to a cellular property.

Theorem 1 (Outstar Pattern Learning)

Suppose that

(I) the functionals are chosen to keep system trajectories bounded;

(II) a local flow condition holds:

$$\int_0^\infty B(t)dt = \infty \quad \text{only if} \quad \int_0^\infty D(t)dt = \infty; \tag{28}$$

(III) the UCS is practiced sufficiently often, and there exist positive constants K_1 and K_2 such that for all $T \geq 0$,

$$f(T, T+t) \geq K_1 \quad \text{if} \quad t \geq K_2 \tag{29}$$

where

$$f(U,V) = \int_U^V I(\xi) \exp\left[\int_\xi^V A(\eta)d\eta\right] d\xi. \tag{30}$$

Then, given arbitrary continuous and nonnegative initial data in $t \leq 0$ such that $\sum_j z_j(0) > 0$,

(A) practice makes perfect:

The LTM ratios $Z_i(t)$ are monotonically attracted to the UCS weights θ_i if

$$[Z_i(0) - X_i(0)][X_i(0) - \theta_i] \geq 0, \tag{31}$$

or may oscillate at most once due to prior learning if (31) does not hold, no matter how wildly A, B, C, D, and I oscillate;

(B) the UCS is registered in STM and partial learning occurs:

The limits $Q_i = \lim_{t\to\infty} X_i(t)$ and $P_i = \lim_{t\to\infty} Z_i(t)$ exist with

$$Q_i = \theta_i, \quad \text{for all } i. \tag{32}$$

(C) If, moreover, the CS is practiced sufficiently often, then perfect learning occurs:

$$\text{if } \int_0^\infty D(t)dt = \infty, \quad \text{then} \quad P_i = \theta_i, \quad \text{for all } i \tag{33}$$

Remarkably, similar global theorems hold for systems (20)–(21) wherein arbitrarily many sampling cells can be simultaneously active and mutually signal each other by complex feedback rules (Geman, 1981; Grossberg, 1969d, 1971a, 1972a, 1980b). This is because all systems of the form (20)–(21) can *factorize* information about how STM and LTM pattern variables learn pattern θ_i from information about how fast energy $I_i(t)$ is being pumped into the system to drive the learning process. The pattern variables Z_i therefore oscillate at most once even if wild fluctuations in input and feedback signal

energies occur through time. In the best theorems now available, only one hypothesis is not known to be necessary and sufficient (Grossberg, 1972a, 1982a).

When many sampling cells v_j, can send sampling signals to each sampled cell v_i, the outstar property that each relative LTM trace $Z_{ji} = z_{ji}(\sum_{k \in I} z_{jk})^{-1}$ oscillates at most once fails to hold. This is so because the Z_{ji} of all active nodes v_j track $X_i = x_i(\sum_k x_k)^{-1}$, while X_i tracks θ_i and the Z_{ji} of all active nodes v_j. The oscillations of the functions $Y_i = \max\{Z_{ji} : j \in J\}$ and $y_i = \min\{Z_{ji} : j \in J\}$ can, however, be classified much as the oscillations of each Z_i can be classified in the outstar case. Since each Z_{ji} depends on all z_{jk} for variable k, each Y_i and y_i depends on all z_{jk} for variable j and k. Since also each X_i depends on all x_k for variable k, the learning at each v_i is influenced by *all* x_k and z_{jk}. No single cell analysis can provide an adequate insight into the dynamics of this associative learning process. The main computational properties emerge through interactions on the network level.

Because the oscillations of all X_i, Y_i, and y_i relative to θ_i can be classified, the following generalization of the outstar learning theorem holds.

Theorem 2 (Unbiased Parallel Pattern Learning)

Suppose that

(I) the functionals are chosen to keep system trajectories bounded;

(II) every sampling cell obeys a local flow condition:

$$\text{for every } j, \quad \int_0^\infty B_j dt = \infty \quad \text{only if} \quad \int_0^\infty D_j dt = \infty; \tag{34}$$

(III) the UCS is presented sufficiently often:

There exist positive constants K_1 and K_2 such that (29) holds.

Then given arbitrary nonnegative and continuous initial data in $t \leq 0$ such that $\sum_i x_i(0) > 0$ and all $\sum_i z_{ji}(0) > 0$,

(A) the UCS is registered in STM and partial learning occurs:

The limits $Q_i = \lim_{t \to \infty} X_i(t)$ and $P_{ji} = \lim_{t \to \infty} Z_{ji}(t)$ exist with

$$Q_i = \theta_i, \quad \text{for all } i. \tag{35}$$

(B) If the jth CS is practiced sufficiently often, then it learns the UCS pattern perfectly:

$$\text{if } \int_0^\infty D_j dt = \infty \quad \text{then} \quad P_{ji} = \theta_i, \text{ for all } i. \tag{36}$$

Because LTM traces z_{ji} gate the performance signals B_j which are activated by a retrieval probe r_j, the theorem enables any and all nodes v_j which sampled the pattern θ_i during learning trials to read it out accurately on recall trials. The theorem does not deny that oscillations in overall network activity can occur during learning and recall, but shows that these oscillations merely influence the rates and intensities of learning and recall. In particular, phase transitions in memory can occur, and the nature of the phases can depend on a complex interaction between network rates and geometry (Grossberg, 1969g, 1982a).

Neither Theorem 1 nor Theorem 2 assumes that the CS and UCS are presented at correlated times. This is because the UCS condition keeps the baseline STM activity of sampled cells from ever decaying below the positive value K_1 in (29). For purposes of space-time pattern learning, this UCS uniformity condition is too strong. In Grossberg (1972a), I used a weaker condition which guarantees that CS–UCS presentations are well enough correlated to guarantee perfect pattern learning of a given spatial pattern by certain cells v_j, even if other spatial patterns are presented at irregular times when they are sampled by distinct cells v_k.

14. PATTERN CALCULUS: RETINA, COMMAND CELL, REWARD, ATTENTION, MOTOR SYNERGY

Three simple but fundamental facts emerge from the mathematical analysis of pattern learning: the unit of LTM is a spatial pattern $\theta = (\theta_i : i \in I)$; suitably designed neural networks can factorize invariant pattern θ from fluctuating energy; the size of a node's sampling signal can render it adaptively sensitive or blind to a pattern θ. These concepts helped me to think in terms of pattern transformations, rather than in terms of feature detectors, computer programs, linear systems, or other types of analysis. When I confronted data about other behavioral problems with these pattern processing properties, the conceptual pressure that was generated drove me into a wide-ranging series of specialized investigations.

What is the minimal network that can discriminate θ from background input fluctuations? It looks like a retina, and the θ's became reflectances (Grossberg, 1970a, 1972b, 1976b, 1983). What is the minimal network that can encode and/or perform a space-time pattern or ordered series of spatial patterns? Called an avalanche, it looks like an invertebrate command cell (Grossberg, 1969e, 1970b). How can one synchronize CS–UCS sampling if the time intervals between CS and UCS presentations are unsynchronized? This analysis led to psychophysiological mechanisms of reward, punishment, and attention (Grossberg, 1971b, 1972c, 1972d, 1975). What are the associative invariants of motor learning? Spatial patterns become motor synergies wherein fixed relative contraction rates across muscles occur, and temporally synchronized performance signals read-out the synergy as a unit (Grossberg, 1970a, 1974).

15. SHUNTING COMPETITIVE NETWORKS OR ADDITIVE NETWORKS?

These specialized investigations repeatedly led to consideration of competitive systems. For example, the same competitive normalization property that arose during modeling of receptor-bipolar-horizontal cell interactions in retina (Grossberg, 1970a, 1972b) also arose in studies of the decision rules needed to release the right amount of incentive motivation in response to interacting drives and conditioned reinforcer inputs within midbrain reinforcement centers (Grossberg, 1972c, 1972d). Because these problems were approached from a behavioral perspective, I knew what interactive properties the competition had to have. I typically found that shunting competition had all the properties that I needed, whereas additive competition often did not. Additive networks approximate shunting networks when their activities are far from cell saturation levels (B_i and D_i in equation (3)). When this is not the case, the automatic gain control properties of shunting networks play a major role, as the next section shows.

16. THE NOISE–SATURATION DILEMMA: PATTERN PROCESSING BY COMPETITIVE NETWORKS

One basic property that was shared by all these systems concerned the manner in which cellular tissues process input patterns whose amplitudes may fluctuate over a much wider range than the cellular activations themselves. This theme is invisible to theories based on binary codes, feature detectors, or additive models. All cellular systems need to prevent sensitivity loss in their responses to both low and high input intensities. Mass action, or shunting, competition enables cells to elegantly solve this problem using automatic gain control by lateral inhibitory signals (Grossberg, 1970a, 1970b, 1973, 1980a). Additive competition fails in this task because it does not, by definition, possess an automatic gain control property.

Suppose that the STM traces or activations $x_1, x_2, \ldots, x_n$ at a network level fluctuate within fixed finite limits at their respective network nodes, as in (3). Setting a bounded operating range for each x_i enables fixed decision criteria, such as output thresholds, to be defined. On the other hand, if a large number of intermittent input sources converge on the nodes through time, then a serious design problem arises, due to the fact that the total input converging on each node can vary wildly through time. I have called this problem the *noise-saturation dilemma* : If the x_i are sensitive to large inputs, then why do not small inputs get lost in internal system noise? If the x_i are sensitive to small inputs, then why do they not all saturate at their maximum values in response to large inputs? Shunting cooperative-competitive networks possess automatic gain control properties capable of generating an infinite dynamic range within which input patterns can be effectively processed, thereby solving the noise-saturation dilemma. The simplest feedforward network will be described to illustrate how its solves the sensitivity problem raised by the noise-saturation dilemma.

Let a spatial pattern $I_i = \theta_i I$ of inputs be processed by the cells v_i, $i = 1, 2, \ldots, n$. Each θ_i is the constant relative size, or reflectance, of its input I_i and I is the variable total input size. In other words, $I = \sum_{k=1}^{n} I_k$, so that $\sum_{k=1}^{n} \theta_k = 1$. How can each cell v_i maintain its sensitivity to θ_i when I is parametrically increased? How is saturation avoided? To compute $\theta_i = I_i(\sum_{k=1}^{n} I_k)^{-1}$, each cell v_i must have information about all the inputs I_k, $k = 1, 2, \ldots, n$. Moreover, since $\theta_i = I_i(I_i + \sum_{k \neq i} I_k)^{-1}$, increasing I_i increases θ_i whereas increasing any I_k, $k \neq i$, decreases θ_i. When this observation is translated into an anatomy for delivering feedforward inputs to the cells v_i, it suggests that I_i excites v_i and that all I_k, $k \neq i$, inhibit v_i. This rule represents the simplest feedforward on-center off-surround anatomy.

How does the on-center off-surround anatomy activate and inhibit the cells v_i via mass action? Let each v_i possess B excitable sites of which $x_i(t)$ are excited and $B - x_i(t)$ are unexcited at each time t. Then at v_i, I_i excites $B - x_i$ unexcited sites by mass action, and the total inhibitory input $\sum_{k \neq i} I_k$ inhibits x_i excited sites by mass action. Moreover, excitation x_i can spontaneously decay at a fixed rate A, so that the cell can return to an equilibrium point (arbitrarily set equal to 0) after all inputs cease. These rules say that

$$\frac{d}{dt} x_i = -A x_i + (B - x_i) I_i - x_i \sum_{k \neq i} I_k. \tag{37}$$

Equation (37) is perhaps the simplest example that illustrates the utility of shunting networks (3). If a fixed spatial pattern $I_i = \theta_i I$ is presented and the background input I

is held constant for awhile, each x_i approaches an equilibrium value. This value is easily found by setting $dx_i/dt = 0$ in (37). It is

$$x_i = \theta_i \frac{BI}{A+I}. \tag{38}$$

Equation (38) represents another example of the factorization of pattern (θ_i) and energy ($BI(A+I)^{-1}$). As a result, the relative activity $X_i = x_i(\sum_{k=1}^{n} x_k)^{-1}$ equals θ_i no matter how large I is chosen; there is no saturation. This is due to automatic gain control by the inhibitory inputs. In other words, $\sum_{k \neq i} I_k$ multiplies x_i in (37). The total gain in (37) is found by writing

$$\frac{d}{dt}x_i = -(A+I)x_i + BI_i. \tag{39}$$

The gain is the coefficient of x_i, namely $-(A+I)$, since if $x_i(0) = 0$,

$$x_i(t) = \theta_i \frac{BI}{A+I}(1 - e^{-(A+I)t}). \tag{40}$$

Both the steady state and the gain of x_i depend on the input strength. This is characteristic of shunting networks but not of additive networks.

The simple law (38) combines two types of information: information about pattern θ_i, or "reflectances", and information about background activity, or "luminance". In visual psychophysics, the tendency towards reflectance processing helps to explain brightness constancy (Grossberg and Todorović, 1988). Another property of (38) is that the total activity

$$x = \sum_{k=1}^{n} x_k = \frac{BI}{A+I} \tag{41}$$

is independent of the number of active cells. This *normalization* rule is a conservation law which says, for example, that a network that receives a fixed total luminance, making one part of the field brighter tends to make another part of the field darker. This property helps to explain brightness contrast, as well as brightness assimilation and the Craik-O'Brien-Cornsweet effect (Grossberg and Todorović, 1988).

Equation (38) can be written in another form that expresses a different physical intuition. If we plot the intensity of an on-center input in logarithmic coordinates K_i, then $K_i = ln(I_i)$ and $I_i = \exp(K_i)$. Also write the total off-surround input as $J_i = \sum_{k \neq i} I_k$. Then (38) can be written in logarithmic coordinates as

$$x_i(K_i, J_i) = \frac{Be^{K_i}}{A + e^{K_i} + J_i}. \tag{42}$$

How does the response x_i at v_i change if we parametrically change the off-surround input J_i? The answer is that x_i's entire response curve to K_i is shifted. Its range of maximal sensitivity scales the off-surround intensity, but its dynamic range is not compressed. Such a shift occurs, for example, in the Weber-Fechner law (Cornsweet, 1970), in bipolar cells of the *Necturus* retina (Werblin, 1971) and in a modified form in the psychoacoustic data of Iverson and Pavel (1981). The shift property says that

$$x_i(K_i + S, J_i^{(1)}) = x_i(K_i, J_i^{(2)}) \tag{43}$$

for all $K_i \geq 0$, where the amount of shift S caused by changing the total off-surround input from $J_i^{(1)}$ to $J_i^{(2)}$ is predicted to be

$$S = ln\left(\frac{A + J_i^{(1)}}{A + J_i^{(2)}}\right). \tag{44}$$

Equation (37) is a special case of a law that occurs *in vivo*; namely, the membrane equation on which cellular neurophysiology is based. The membrane equation describes the voltage $V(t)$ of a cell by the law

$$C\frac{\partial V}{\partial t} = (V^+ - V)g^+ + (V^- - V)g^- + (V^p - V)g^p. \tag{45}$$

In (45), C is a capacitance; V^+, V^-, and V^p are constant excitatory, inhibitory, and passive saturation points, respectively; and g^+, g^-, and g^p are excitatory, inhibitory, and passive conductances, respectively. We will scale V^+ and V^- so that $V^+ > V^-$. Then *in vivo* $V^+ \geq V(t) \geq V^-$ and $V^+ > V^p \geq V^-$. Often V^+ represents the saturation point of a Na^+ channel and V^- represents the saturation point of a K^+ channel. To see why (37) is a special case of (45), suppose that (45) holds at each cell v_i. Then at v_i, $V = x_i$. Set $C = 1$ (rescale time), $V^+ = B$, $V^- = V^p = 0$, $g^+ = I_i$, $g^- = \sum_{k \neq i} I_k$, and $g^p = A$.

There is also symmetry-breaking in (45) because $V^+ - V^p$ is usually much larger than $V^p - V^-$. This symmetry-breaking operation, which is usually mentioned in the experimental literature without comment, achieves an important noise-suppression property when it is coupled to an on-center off-surround anatomy. For example, in the network

$$\frac{d}{dt}x_i = -Ax_i + (B - x_i)I_i - (x_i + C)\sum_{k \neq i} I_k, \tag{46}$$

both depolarized potentials ($0 < x_i \leq B$) and hyperpolarized potentials ($-C \leq x_i < 0$) can occur. The equilibrium activity in response to spatial pattern $I_i = \theta_i I$ is

$$x_i = \frac{(B + C)I}{A + I}[\theta_i - \frac{C}{B + C}]. \tag{47}$$

Parameter $C(B + C)^{-1}$ is an *adaptation level* which θ_i must exceed in order to depolarize x_i and thereby generate an output signal from v_i. In order to inhibit uniform input patterns that do not carry discriminative featural information, we would want $\theta_i = \frac{1}{n}$ for all i to imply that all $x_i = 0$. This occurs if $B = (n - 1)C$, so that $B \gg C$ and thus $V^+ - V^p \gg V^p - V^-$.

The reflectance processing and Weber law properties, the total activity normalization property, and the adaptation level property of (46) set the stage for the design and classification of more complex feedforward and feedback on-center off-surround shunting networks during the early 1970's.

17. SHORT TERM MEMORY STORAGE AND CAM

Feedback networks are capable of storing memories in STM for far longer than a passive decay rate, such as A in (37), would allow, yet can also be rapidly reset. The

simplest feedback competitive network capable of solving the noise-saturation dilemma is defined by the equations

$$\frac{d}{dt}x_i = -Ax_i + (B - x_i)[I_i + f(x_i)] - x_i[J_i + \sum_{k \neq i} f(x_k)], \tag{48}$$

$i = 1, 2, \ldots, n$. Suppose that the inputs I_i and J_i acting before $t = 0$ establish an arbitrary initial activity pattern $(x_1(0), x_2(0), \ldots, x_n(0))$ before being shut off at $t = 0$. How does the choice of the feedback signal function $f(w)$ control the transformation and storage of this pattern as $t \to \infty$?

The answer is determined by the choice of function $g(w) = w^{-1}f(w)$, which measures how much $f(w)$ deviates from linearity at prescribed activity levels w. The network's responses to these choices may be summarized using the functions $X_i = x_i(\sum_{k=1}^n x_k)^{-1}$ and $x = \sum_{k=1}^n x_k$. The relative activity X_i of the ith node computes how the network transforms the input pattern through time. The functions X_i play the role for feedback networks that the reflectances θ_i in (38) play for feedforward networks; also recall Theorems 1 and 2. The total activity x measures how well the network normalizes the total network activity and whether the pattern is stored ($x(\infty) = \lim_{t \to \infty} x(t) > 0$) or not ($x(\infty) = 0$). Variable x plays the role of the total input I in (38). In Grossberg (1973) the following types of results were proved about these systems: Linear signals lead to perfect pattern storage and noise amplification. Slower-than-linear signals lead to pattern uniformization and noise amplification. Faster-than-linear signals lead to winner-take-all choices, noise suppression, and total activity quantization in a network that behaves like an emergent finite state machine. Sigmoid signals lead to partial contrast enhancement, tunable filtering, noise suppression, and normalization. See Grossberg (1981, 1988) for reviews.

All of these networks function as a type of global content addressable memory, or CAM, since all trajectories converge to equilibrium points through time. The equilibrium point to which the network converges in response to an input pattern plays the role of a stored memory. Both linear and sigmoid signals can be chosen to create networks with infinitely many, indeed nondenumerably many, equilibria. Faster-than-linear signals give rise to only finitely many equilibria as part of their winner-take-all property.

In summary, several factors work together to generate desirable pattern transformation and STM storage properties. The dynamics of mass action, the geometry of competition, and the statistics of competitive feedback signals work together to define a unified network module whose several parts are designed in a coordinated fashion through development.

18. EVERY COMPETITIVE SYSTEM INDUCES A DECISION SCHEME

As solutions to specialized problems involving competition accumulated, networks capable of normalization, sensitivity changes via automatic gain control, attentional biases, developmental biases, pattern matching, shift properties, contrast enhancement, edge and curvature detection, tunable filtering, multistable choice behavior, normative drifts, traveling waves, synchronous oscillations, hysteresis, and resonance began to be classified within the framework of additive or shunting feedforward or feedback competitive networks. As in the case of associative learning, the abundance of special cases made it

seem more and more imperative to find a mathematical framework within which these results could be unified and generalized. I also began to realize that many of the pattern transformations and STM storage properties of specialized examples were instances of an absolute stability property of a general class of networks.

This unifying mathematical theme can be summarized intuitively as follows: every competitive system induces a decision scheme that can be used to prove global limit and oscillation theorems, notably absolute stability theorems (Grossberg, 1978c, 1978d, 1980c). This decision scheme interpretation provides a geometrical way to think about a Liapunov functional that is naturally associated with every competitive system. A competitive dynamical system is, for present purposes, defined by a system of differential equations such that

$$\frac{d}{dt}x_i = f_i(x_1, x_2, \quad , x_n) \tag{49}$$

where

$$\frac{\partial f_i}{\partial x_j} \leq 0, \quad i \neq j, \tag{50}$$

and the f_i are chosen to generate bounded trajectories. By (50), increasing the activity x_j of a given population can only decrease the growth rates $\frac{d}{dt}x_i$ of other populations, $i \neq j$, or may not influence them at all. No constraint is placed upon the sign of $\frac{\partial f_i}{\partial x_i}$. Typically, cooperative behavior occurs within a population and competitive behavior occurs between populations, as in the on-center off-surround networks (48).

The method makes mathematically precise the intuitive idea that a competitive system can be understood by keeping track of who is winning the competition. The decision scheme makes this intuition precise. To define it, write (49) in the form

$$\frac{d}{dt}x_i = a_i(x_i)M_i(x), \quad x = (x_1, x_2, \quad , x_n)$$

which factors out the amplification function $a_i(x_i) \geq 0$. Then define

$$M^+(x) = \max\{M_i(x) : i = 1, 2, \ldots, n\}$$

and

$$M^-(x) = \min\{M_i(x) : i = 1, 2, \ldots, n\}. \tag{53}$$

These variables track the largest and smallest rates of change, and are used to keep track of who is winning. Using these functions, it is easy to see that there exists a property of *ignition*: Once a trajectory enters the *positive ignition region*

$$R^+ = \{x : M^+(x) \geq 0\} \tag{54}$$

or the *negative ignition region*

$$R^- = \{x : M^-(x) \leq 0\},$$

it can never leave it. If $x(t)$ never enters the set

$$R^* = R^+ \cap R^-, \tag{56}$$

then each variable $x_i(t)$ converges monotonically to a limit. The interesting behavior in a competitive system occurs in R^*. In particular, if $x(t)$ never enters R^+, each $x_i(t)$ decreases to a limit; then the competition never gets started. The set

$$S^+ = \{x : M^+(x) = 0\} \tag{57}$$

acts like a competition threshold, which is called the *positive ignition hypersurface.*

We therefore consider a trajectory after it has entered R^*. For simplicity, redefine the time scale so that the trajectory is in R^* at time $t = 0$. The Liapunov functional for any competitive system is then defined as

$$L(x_t) = \int_0^t M^+(x(v))dv.$$

The Liapunov property is a direct consequence of positive ignition:

$$\frac{d}{dt}L(x_t) = M^+(x(t)) \geq 0. \tag{59}$$

This functional provides the "energy" that forces trajectories through a series of competitive decisions, which are also called *jumps*. Jumps keep track of the state which is undergoing the *maximal* rate of change at any time ("who's winning"). If $M^+(x(t)) = M_i(x(t))$ for times $S \leq t < T$ but $M^+(x(t)) = M_j(x(t))$ for times $T \leq t < U$, then we say that the system *jumps* from node v_i to node v_j at time $t = T$. A jump from v_i to v_j can only occur on the *jump set*

$$J_{ij} = \{x \in R^* : M^+(x) = M_i(x) = M_j(x)\}. \tag{60}$$

The Liapunov functional $L(x_t)$ moves the system through these decision hypersurfaces through time. The geometry of S^+, S^-, and the jump sets J_{ij}, together with the energy defined by $L(x_t)$, can be used to globally analyse the dynamics of the system. In particular, due to the positive ignition property (59), the limit

$$\lim_{t\to\infty} L(x_t) = \int_0^\infty M^+(x(v))dv \tag{61}$$

always exists, and is possibly infinite.

19. LIMITS AND OSCILLATIONS: CONSENSUS AND CONTRADICTION

The following results illustrate the use of these concepts (Grossberg, 1978c):

Theorem 3: Given any initial data $x(0)$, suppose that

$$\int_0^\infty M^+(x(v))dv < \infty. \tag{62}$$

Then the limit $x(\infty) = \lim_{t\to\infty} x(t)$ exists.

Corollary 1: If in response to initial data $x(0)$, all jumps cease after some time $T < \infty$, then $x(\infty)$ exists.

Speaking intuitively, this result means that after all local decisions, or jumps, have been made in response to an initial state $x(0)$, then the system can settle down to a

global decision, or equilibrium point $x(\infty)$. In particular, if $x(0)$ leads to only finitely many jumps because there exists a jump tree, or partial ordering of decisions, then $x(\infty)$ exists. This fact led to the analysis of circumstances under which no jump cycle, or repetitive series of jumps, occurs in response to $x(0)$, and hence that jump trees exist. These results included examples of nonlinear Volterra-Lotka equations with asymmetric interaction equations all of whose trajectories approach equilibrium points (Grossberg, 1978c). Thus symmetric coefficients were shown not to be necessary for global approach to equilibrium, or a global CAM property, to obtain.

Further information may be derived from (62). Since $M^+(x(t)) \geq 0$ for all $t \geq 0$, it also follows that $\lim_{t\to\infty} M^+(x(t)) = 0$. This tells us to look for the equilibrium points $x(\infty)$ on the positive ignition hypersurface S^+ in (57):

Corollary 2: If $\int_0^\infty M^+(x(t))dt < \infty$, then $x(\infty) \in S^+$.

Thus the positive ignition surface is the place where the competition both ignites and its memories are stored if no jump cycle exists. Using this result, an analysis was made of conditions under which no jump cycle exists in response to any initial vector $x(0)$, and hence all trajectories approach an equilibrium state.

The same method was also used to prove that a competitive system can generate sustained oscillations if it contains globally inconsistent decisions. These results provide examples where asymmetric coefficients do lead to oscillations. Here, in response to initial data $x(0)$,

$$\int_0^\infty M^+(x(v))dv = \infty, \tag{63}$$

thus infinitely many jumps occur, hence a jump cycle occurs, and the trajectory undergoes undamped oscillations. This method was used to provide a global analysis of the oscillations taking place in a variety of competitive systems, including the Volterra-Lotka systems that model the voting paradox (Grossberg, 1978c, 1980c; May and Leonard, 1975).

Using this method, a large new class of nonlinear competitive networks was identified all of whose trajectories converge to one of possibly infinitely many equilibrium points (Grossberg, 1978d). These are the *adaptation level systems*

$$\frac{d}{dt}x_i = a_i(x)[b_i(x_i) - c(x)] \tag{64}$$

which were identified through an analysis of many specialized networks. In system (63), each state-dependent amplification function $a_i(x)$ and self-signal function $b_i(x_i)$ can be chosen with great generality without destroying the system's ability to reach equilibrium because there exists a state-dependent *adaptation level* $c(x)$ against which each $b_i(x_i)$ is compared. Such an adaptation level $c(x)$ defines a strong type of long-range symmetry within the system. Equation (64) is a feedback analog of the feedforward adaptation level equation (47).

The examples which motivated the analysis of (64) were additive networks

$$\frac{d}{dt}x_i = -A_i x_i + \sum_k f_k(x_k)B_{ki} + I_i$$

and shunting networks

$$\frac{d}{dt}x_i = -A_i x_i + (B_i - x_i)[I_i + \sum_k f_k(x_k)C_{ki}] - (x_i + D_i)[J_i + \sum_k g_k(x_k)E_{ki}] \tag{66}$$

in which the symmetric coefficients B_{ki}, C_{ki}, and E_{ki} took on different values when $k = i$ and when $k \neq i$. Examples in which the symmetric coefficients varied with $| k - i |$ in a graded fashion were also studied through computer simulations (Ellias and Grossberg, 1975; Levine and Grossberg, 1976). An adequate global mathematical convergence proof was announced in Grossberg (1982b) and elaborated in Cohen and Grossberg (1983).

A special case of my theorem concerning these adaption level systems is the following.

Theorem 4 (Absolute Stability of Adaptation Level Systems)

Suppose that

(I) *Smoothness*:

The functions $a_i(x)$, $b_i(x_i)$, and $c(x)$ are continuously differentiable;

(II) *Positivity*:

$$a_i(x) > 0 \quad \text{if} \quad x_i > 0, \quad x_j \geq 0, \quad j \neq i; \tag{67}$$

$$a_i(x) = 0 \quad \text{if} \quad x_i = 0, \quad x_j \geq 0, \quad j \neq i; \tag{68}$$

for sufficiently small $\lambda > 0$, there exists a continuous function $\bar{a}_i(x_i)$ such that

$$\bar{a}_i(x_i) \geq a_i(x) \quad \text{if} \quad x \in [0, \lambda]^n \tag{69}$$

and

$$\int_0^\lambda \frac{dw}{\bar{a}_i(w)} = \infty; \tag{70}$$

(III) *Boundedness*: for each $i = 1, 2, \ldots, n$,

$$\limsup_{x_i \to \infty} b_i(x_i) < c(0, 0, \quad, \infty, 0, \quad, 0) \tag{71}$$

where ∞ is in the ith entry of $(0, 0, \quad, \infty, 0, \quad, 0)$;

(IV) *Competition*:

$$\frac{\partial c(x)}{\partial x_i} > 0, \quad x \in \mathbf{R}^n_+, \quad i = 1, 2, \quad, n; \tag{72}$$

(V) *Decision Hills*:

The graph of each $b_i(x_i)$ possesses at most finitely many maxima in every compact interval.

Then the pattern transformation is stored in STM because all trajectories converge to equilibrium points; that is, given any $x(0) > 0$, the limit $x(\infty) = \lim_{t \to \infty} x(t)$ exists.

This theorem intuitively means that the decision schemes of adaptation level systems are globally consistent and give rise to a global CAM.

In the proof of Theorem 4, it was shown that each $x_i(t)$ gets trapped within a sequence of decision boundaries that get laid down through time at the abscissa values of the highest peaks in the graphs of the functions b_i in (64). The size and location of these peaks reflect the statistical rules, which can be chosen extremely complex, that give rise to the output signals from the totality of cooperating subpopulations within each node v_i. In particular, a b_i with multiple peaks can be generated when a population's positive feedback signal function is a multiple-sigmoid function which adds up output signals from multiple randomly defined subpopulations within v_i. After all the decision boundaries get laid down, each x_i is trapped within a single valley of its b_i graph. This valley acts, in some respects, like a classical potential. After all the x_i get trapped in such valleys, the function

$$B[x(t)] = \max\{b_i(x(t)) : i = 1, 2, \ldots, n\} \tag{73}$$

is a Liapunov function. This Liapunov property was used to complete the proof of the theorem.

Adaptation level systems exclude distance-dependent interactions. To overcome this gap, Michael Cohen and I (Cohen and Grossberg, 1983; see also Grossberg, 1982b) studied the absolute stability of the symmetric networks

$$\frac{d}{dt}x_i = -A_i x_i + (B_i - C_i x_i)[I_i + f_i(x_i)] - (D_i x_i + E_i)[J_i + \sum_{k=1} g_k(x_k)F_{ki}], \tag{74}$$

where $F_{ij} = F_{ji}$. The adaptation level model (64) is in some ways more general and in some ways less general than model (74). Cohen and I began our study of (74) with the hope that we could use the symmetric coefficients in (74) to prove that no jump cycles exist, and thus that all trajectories approach equilibrium as a consequence of Theorem 3. Such a proof would be part of a more general theory and, by using geometrical concepts such as jump set and ignition surface, it would clarify how to perturb off the symmetric coefficients without generating oscillations.

As it turned out, the global Liapunov method that I developed in the 1970's sensitized us to think in that direction. We soon discovered a general class of symmetric models and a global Liapunov function for every model in the class. In each of these models, the Liapunov function was used to prove that all trajectories approach equilibrium points. This CAM model, which is now often called the Cohen-Grossberg model, was designed to include additive networks (65) and shunting networks (66) with symmetric coefficients.

20. COHEN–GROSSBERG CAM MODEL AND THEOREM

The Cohen-Grossberg model includes any dynamical system that can be written in the form

$$\frac{d}{dt}x_i = a_i(x_i)[b_i(x_i) - \sum_{j=1}^{n} c_{ij}d_j(x_j)] \tag{75}$$

Each such model admits the global Liapunov function

$$V = -\sum_{i=1}^{n}\int^{x_i} b_i(\xi_i)d_i'(\xi_i)d\xi_i + \frac{1}{2}\sum_{j,k=1}^{n} c_{jk}d_j(x_j)d_k(x_k) \tag{76}$$

if the coefficient matrix $C = \| c_{ij} \|$ and the functions a_i, b_i, and d_j obey mild technical conditions, including

Symmetry:

$$c_{ij} = c_{ji}, \tag{77}$$

Positivity:

$$a_i(x_i) \geq 0 \tag{78}$$

Monotonicity:

$$d_j'(x_j) \geq 0. \tag{79}$$

Integrating V along trajectories implies that

$$\frac{d}{dt}V = -\sum_{i=1}^{n} a_i d_i'[b_i - \sum_{j=1}^{n} c_{ij} d_j]^2. \tag{80}$$

If (78) and (79) hold, then $\frac{d}{dt}V \leq 0$ along trajectories. Once this basic property of a Liapunov function is in place, it is a technical matter to rigorously prove that every trajectory approaches one of a possibly large number of equilibrium points.

For expository vividness, the functions in the Cohen-Grossberg model (75) are called the *amplification* function a_i, the *self-signal* function b_i, and the *other-signal* functions d_j. Specialized models are characterized by particular choices of these functions.

A. Additive Model

Cohen and Grossberg (1983, p.819) noted that "the simpler additive neural networks ... are also included in our analysis". The additive equation (2) can be written using the coefficients of the standard electrical circuit interpretation (Plonsey and Fleming, 1969) as

$$C_i \frac{dx_i}{dt} = -\frac{1}{R_i}x_i + \sum_{j=1}^{n} f_j(x_j) z_{ji} + I_i. \tag{81}$$

Substitution into (75) shows that

$$a_i(x_i) = \frac{1}{C_i} \quad \text{(constant!)} \tag{82}$$

$$b_i(x_i) = \frac{1}{R_i}x_i + I_i \quad \text{(linear!)} \tag{83}$$

$$c_{ij} = -T_{ij} \tag{84}$$

and

$$d_j(x_j) = f_j(x_j). \tag{85}$$

Thus in the additive case, the amplification function (82) is a positive constant, hence satisfies (78), and the self-signal term (83) is linear. Substitution of (82)–(83) into (76) leads directly to the equation

$$V = \sum_{i=1}^{n} \frac{1}{R_i} \int^{x_i} \xi_i f_i'(\xi_i) d\xi_i - \sum_{i=1}^{n} I_i f_i(x_i) - \frac{1}{2} \sum_{j,k=1}^{n} T_{jk} f_j(x_j) f_k(x_k). \tag{86}$$

This Liapunov function for the additive model was later published by Hopfield (1984). In Hopfield's treatment, ξ_i is written as an inverse $f_i^{-1}(V_i)$. Cohen and Grossberg (1983) showed, however, that although $f_i(x_i)$ must be nondecreasing, as in (79), it need not have an inverse in order for (86) to be valid.

B. Shunting Model

All additive models lead to constant amplification functions $a_i(x_i)$ and linear self-feedback functions $b_i(x_i)$. The need for the more general model (75) becomes apparent when the shunting STM equation (3) is analysed. Consider, for example, a class of shunting models.

$$\frac{d}{dt}x_i = -A_i x_i + (B_i - x_i)[I_i + f_i(x_i)] - (x_i + C_i)[J_i + \sum_{j=1}^{n} D_{ij} g_j(x_j)]. \tag{87}$$

In (87), each x_i can fluctuate within the finite interval $[-C_i, B_i]$ in response to the constant inputs I_i and J_i, the state-dependent positive feedback signal $f_i(x_i)$, and the negative feedback signals $D_{ij}g_j(x_j)$. It is assumed that

$$D_{ij} = D_{ji} \geq 0 \tag{88}$$

and that

$$g_j'(x_j) \geq 0. \tag{89}$$

In order to write (87) in Cohen-Grossberg form, it is convenient to introduce the variables

$$y_i = x_i + C_i. \tag{90}$$

In applications, C_i is typically nonnegative. Since x_i can vary within the interval $[-C_i, B_i]$, y_i can vary within the interval $[0, B_i + C_i]$ of nonnegative numbers. In terms of these variables, (87) can be written in the form

$$\frac{d}{dt}y_i = a_i(y_i)[b_i(y_i) - \sum_{j=1}^{n} C_{ij} d_j(y_j)] \tag{91}$$

where

$$a_i(y_i) = y_i \quad \text{(nonconstant!)}, \tag{92}$$

$$b_i(y_i) = \frac{1}{x_i}[A_i C_i - (A_i + J_i)x_i + (B_i + C_i - x_i)(I_i + f_i(x_i - C_i))] \quad \text{(nonlinear!)}, \tag{93}$$

$$C_{ij} = D_{ij}, \tag{94}$$

and

$$d_j(y_j) = g_j(y_j - C_j) \quad \text{(noninvertible!)}. \tag{95}$$

Unlike the additive model, the amplification function $a_i(y_i)$ in (92) is not a constant. In addition, the self-signal function $b_i(y_i)$ in (93) is not necessarily linear, notably because the feedback signal $f_i(x_i - C_i)$ is often nonlinear in applications of the shunting model; in particular it is often a sigmoid or multiple sigmoid signal function.

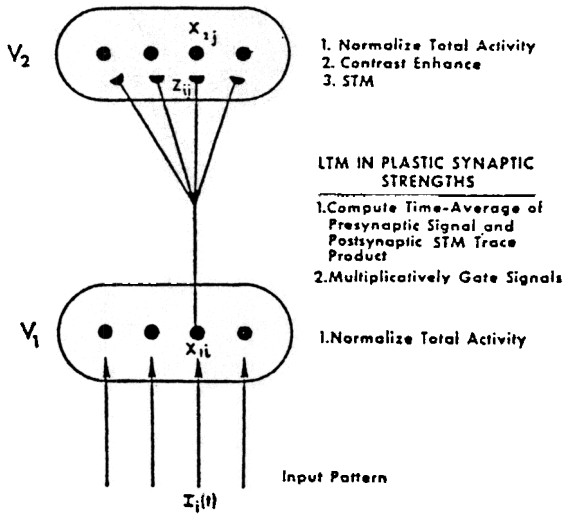


Figure 6. The basic computational rules of self-organizing feature maps were established by 1976. (Reprinted with permission from Grossberg, 1976b.)

Property (78) follows from the fact that $a_i(y_i) = y_i \geq 0$. Property (79) follows from the assumption that the negative feedback signal function g_j is monotone nondecreasing. Cohen and Grossberg (1983) proved that g_j need not be invertible. A signal threshold may exist below which $g_j = 0$ and above which g_j may grow in a nonlinear way. The inclusion of nonlinear signals with thresholds better enables the model to deal with fluctuations due to subthreshold noise.

These results show that adaptation level and distance-dependent competitive networks represent stable neural designs for competitive decision-making and CAM. The fact that adaptation level systems have been analyzed using Liapunov functionals whereas distance-dependent, and more generally, symmetric networks have been analyzed using Liapunov functions shows that the global convergence theory of competitive systems is still incomplete. Global limit theorems for cooperative systems were also subsequently discovered (Hirsch, 1982, 1985, 1989), as were theorems showing when closely related cooperative-competitive systems could oscillate (Cohen, 1988, 1990). Major progress has also been made on explicitly constructing dynamical systems with prescribed sets of equilibrium points, and only these equilibrium points (Cohen, 1992). This is an exciting area for intensive mathemâtical investigation. Additive and shunting networks have also found their way into many applications. Shunting networks have been particularly useful in understanding biological and machine vision, from the earliest retinal detection stages through higher cortical filtering and grouping processes (Gaudiano, 1992a, 1992b: Grossberg and Mingolla, 1985a, 1985b; Nabet and Pinter, 1991), as well as perceptual and motor oscillations (Cohen, Grossberg, and Pribe, 1993; Gaudiano and Grossberg, 1991; Grossberg and Somers, 1991, 1992; Somers and Kopell, 1993).

21. COMPETITIVE LEARNING AND SELF-ORGANIZING FEATURE MAPS

Once mathematical results were available that clarified the global dynamics of associative learning and competition, the stage was set to combine these mechanisms in models of cortical development, recognition learning, and categorization. One major source of interest in such models came from neurobiological experiments on geniculocortical and retinotectal development (Gottlieb, 1976; Hubel and Wiesel, 1977; Hunt and Jacobson, 1974). My own work on this problem was stimulated by such neural data, and by psychological data concerning perception, cognition, and motor control. Major constraints on theory construction also derived from my previous results on associative learning. During outstar learning, for example, no learning of a sampled input pattern θ_i in (27) occurs, $i = 1, 2, \ldots, n$, when the learning signal $D(t) = 0$ in equation (26). This property was called *stimulus sampling*. It showed that activation of an outstar source cell enables it to selectively learn spatial patterns at prescribed times. This observation led to the construction of more complex sampling cells and networks, called *avalanches*, that are capable of learning arbitrary *space–time* patterns, not merely *spatial* patterns, and to a comparison of avalanche networks with properties of command cells in invertebrates (Grossberg, 1969e, 1970b, 1974).

Activation of outstars and avalanches needs to be selective, so as not to release, or recall, learned responses in unappropriate contexts. Networks were needed that could selectively filter input patterns so as to activate outstars and avalanches only under appropriate stimulus conditions. This work led to the introduction of instar networks in Grossberg (1970a, 1972b), to the description of the first self-organizing feature map in Malsburg (1973), and to the development of the main equations and mathematical properties of the modern theory of competitive learning, self-organizing feature maps, and learned vector quantization in Grossberg (1976a, 1976b, 1976c, 1978a). Willshaw and Malsburg (1976) and Malsburg and Willshaw (1977, 1981) also made a seminal contribution at this time to the modelling of cortical development using self-organizing feature maps. In addition, the first self-organizing multilevel networks were constructed in 1976 for the learning of multidimensional maps from $\Re^n$ to $\Re^m$, for any $n, m \geq 1$ (Grossberg, 1976a, 1976b, 1976c). . The first two levels F_1 and F_2 constitute a self-organizing feature map such that input patterns to F_1 are categorized at F_2. Levels F_2 and F_3 are built out of outstars so that categorizing nodes at F_2 can learn output patterns at F_3. Hecht-Nielsen (1987) later called such networks *counterpropagation* networks and claimed that they were a new model. The name instar-outstar map has been used for these maps since the 1970's. Recent popularizers of back propagation have also claimed that multilevel neural networks for adaptive mapping were not available until their work using back propagation in the last half of the 1980's. Actually, back propagation was introduced by Werbos (1974) and self-organizing mapping networks that were proven to be stable in sparse environments were available in 1976. An account of the historical development of self-organizing feature maps is provided in Carpenter and Grossberg (1991).

The main processing levels and properties of self-organizing feature maps are summarized in Figure 6, which is reprinted from Grossberg (1976b). In such a model, an input pattern is normalized and registered as a pattern of activity, or STM, across the feature detectors of level F_1. Each F_1 output signal is multiplied or gated, by the adaptive weight, or LTM trace, in its respective pathway, and all these LTM-gated inputs are added up

at their target F_2 nodes, as in equations (1)–(3). Lateral inhibitory, or competitive, interactions within F_2 contrast-enhance this input pattern; see Section 17. Whereas many F_2 nodes may receive inputs from F_1, lateral inhibition allows a much smaller set of F_2 nodes to store their activation in STM.

Only the F_2 nodes that win the competition and store their activity in STM can influence the learning process. STM activity opens a learning gate at the LTM traces that abut the winning nodes, as in equation (7). These LTM traces can then approach, or track, the input signals in their pathways by a process of steepest descent. This learning law has thus often been called *gated steepest descent*, or *instar learning*. As noted in Section 2, it was introduced into neural network models in the 1960's (e.g. Grossberg, 1969d). Because such an LTM trace can either increase or decrease to track the signals in its pathway, it is not a Hebbian associative law (Hebb, 1949). It has been used to model neurophysiological data about hippocampal LTP (Levy, 1985; Levy and Desmond, 1985) and adaptive tuning of cortical feature detectors during the visual critical period (Rauschecker and Singer, 1979; Singer, 1983), lending support to the 1976 prediction that both systems would employ such a learning law (Grossberg, 1976b, 1978a). Hecht-Nielsen (1987) has called the instar learning law Kohonen learning after Kohonen's use of the law in his applications of self-organizing feature maps in the 1980's, as in Kohonen (1984). The historical development of this law, including its use in self-organizing feature maps in the 1970's, does not support this attribution.

Indeed, after self-organizing feature map models were introduced and computationally characterized in Grossberg (1976b, 1978a), Malsburg (1973), and Willshaw and Malsburg (1976), these models were subsequently applied and specialized by many authors (Amari and Takeuchi, 1978; Bienenstock, Cooper and Munro, 1982; Commons, Grossberg, and Staddon, 1991; Grossberg, 1982a, 1987; Grossberg and Kuperstein, 1986; Kohonen, 1984; Linsker, 1986; Rumelhart and Zipser, 1985). They exhibit many useful properties, especially if not too many input patterns, or clusters of input patterns, perturb level F_1 relative to the number of categorizing nodes in level F_2. It was proved that under these sparse environmental conditions, category learning is stable, with LTM traces that track the statistics of the environment, are self-normalizing, and oscillate a minimum number of times (Grossberg, 1976b, 1978a). Also, the category decision rule, as in a Bayesian classifier, tends to minimize error. It was also proved, however, that under arbitrary environmental conditions, learning becomes unstable. Such a model could forget your parents' faces. Although a gradual switching off of plasticity can partially overcome this problem, such a mechanism cannot work in a recognition learning system whose plasticity is maintained throughout adulthood.

This memory instability is due to basic properties of associative learning and lateral inhibition. An analysis of this instability, together with data about categorization, conditioning, and attention, led to the introduction of Adaptive Resonance Theory, or ART, models that stabilize the memory of self-organizing feature maps in response to an arbitrary stream of input patterns (Grossberg, 1976c). A central prediction of ART, from its inception, has been that adult learning mechanisms share properties with the adaptive mechanisms that control developmental plasticity, in particular that "adult attention is a continuation on a developmental continuum of the mechanisms needed to solve the stability-plasticity dilemma in infants" (Grossberg, 1982b, p. 335). Recent experimental results concerning the neural control of learning have provided increasing support for this

hypothesis (Kandel and O'Dell, 1992).

22. ADAPTIVE RESONANCE THEORY

In an ART model, as shown in Figure 7a, an input vector **I** registers itself as a pattern **X** of activity across level F_1. The F_1 output vector **S** is then transmitted through the multiple converging and diverging adaptive filter pathways emanating from F_1. This transmission event multiplies the vector **S** by a matrix of adaptive weights, or LTM traces, to generate a net input vector **T** to level F_2. The internal competitive dynamics of F_2 contrast-enhance vector **T**. Whereas many F_2 nodes may receive inputs from F_1, competition or lateral inhibition between F_2 nodes allows only a much smaller set of F_2 nodes to store their activation in STM. A compressed activity vector **Y** is thereby generated across F_2. In the ART 1 and ART 2 models (Carpenter and Grossberg, 1987a, 1987b), the competition is tuned so that the F_2 node that receives the maximal $F_1 \rightarrow F_2$ input is selected. Only one component of **Y** is nonzero after this choice takes place. Activation of such a winner-take-all node defines the category, or symbol, of the input pattern **I**. Such a category represents all the inputs **I** that maximally activate the corresponding node. So far, these are the rules of a self-organizing feature map.

In a self-organizing feature map, only the F_2 nodes that win the competition and store their activity in STM can immediately influence the learning process. In an ART model (Carpenter and Grossberg, 1987a, 1992), learning does not occur as soon as some winning F_2 activities are stored in STM. Instead activation of F_2 nodes may be interpreted as "making a hypothesis" about an input **I**. When **Y** is activated, it rapidly generates an output vector **U** that is sent top-down through the second adaptive filter. After multiplication by the adaptive weight matrix of the top-down filter, a net vector **V** inputs to F_1 (Figure 7b). Vector **V** plays the role of a learned top-down expectation. Activation of **V** by **Y** may be interpreted as "testing the hypothesis" **Y**, or "reading out the category prototype" **V**. An ART network is designed to match the "expected prototype" **V** of the category against the active input pattern, or exemplar, **I**. Nodes that are activated by **I** are suppressed if they do not correspond to large LTM traces in the prototype pattern **V**. Thus F_1 features that are not "expected" by **V** are suppressed. Expressed in a different way, the matching process may change the F_1 activity pattern **X** by suppressing activation of all the feature detectors in **I** that are not "confirmed" by hypothesis **Y**. The resultant pattern **X*** encodes the cluster of features in **I** that the network deems relevant to the hypothesis **Y** based upon its past experience. Pattern **X*** encodes the pattern of features to which the network "pays attention."

If the expectation **V** is close enough to the input **I**, then a state of *resonance* develops as the attentional focus takes hold. The pattern **X*** of attended features reactivates hypothesis **Y** which, in turn, reactivates **X***. The network locks into a resonant state through the mutual positive feedback that dynamically links **X*** with **Y**. In ART, the resonant state, rather than bottom-up activation, drives the learning process. The resonant state persists long enough, at a high enough activity level, to activate the slower learning process; hence the term *adaptive resonance* theory. ART systems learn prototypes, rather than exemplars, because the attended feature vector **X***, rather than the input **I** itself, is learned. These prototypes may, however, also be used to encode individual exemplars, as described below.

23. MEMORY STABILITY AND 2/3 RULE MATCHING

This attentive matching process is realized by combining three different types of inputs at level F_1 (Figure 7): bottom-up inputs, top-down expectations, and attentional gain control signals. The attentional gain control channel sends the same signal to all F_1 nodes; it is a "nonspecific", or modulatory, channel. Attentive matching obeys a 2/3 Rule (Carpenter and Grossberg, 1987a): an F_1 node can be fully activated only if two of the three input sources that converge upon it send positive signals at a given time.

The 2/3 Rule allows an ART system to react to bottom-up inputs, since an input directly activates its target F_1 features and indirectly activates them via the nonspecific gain control channel to satisfy the 2/3 Rule (Figure 7a). After the input instates itself at F_1, leading to selection of a hypothesis $\mathbf{Y}$ and a top-down prototype $\mathbf{V}$, the 2/3 Rule ensures that only those F_1 nodes that are confirmed by the top-down prototype can be attended at F_1 after an F_2 category is selected.

The 2/3 Rule enables an ART network to realize a self-stabilizing learning process. Carpenter and Grossberg (1987a) proved that ART learning and memory are stable in arbitrary environments, but become unstable when 2/3 Rule matching is eliminated. Thus a type of matching that guarantees stable learning also enables the network to pay attention.

24. PHONEMIC RESTORATION AND PRIMING

2/3 Rule matching in the brain is illustrated by experiments on phonemic restoration (Repp, 1991; Samuel, 1981a, 1981b; Warren, 1984; Warren and Sherman, 1974). Suppose that a noise spectrum replaces a letter sound in a word heard in an otherwise unambiguous context. Then subjects hear the correct letter sound, not the noise, to the extent that the noise spectrum includes the letter formants. If silence replaces the noise, then only silence is heard. Top-down expectations thus amplify expected input features while suppressing unexpected features, but do not create activations not already in the input.

2/3 Rule matching also shows how an ART system can be primed. This property has been used to explain paradoxical reaction time and error data from priming experiments during lexical decision and letter gap detection tasks (Grossberg and Stone, 1986; Schvaneveldt and MacDonald, 1981). Although priming is often thought of as a residual effect of previous bottom-up activation, a combination of bottom-up activation and top-down 2/3 Rule matching was needed to explain the complete data pattern. This analysis combined bottom-up priming with a type of top-down priming; namely, the top-down activation that prepares a network for an expected event that may or may not occur. The 2/3 Rule clarifies why top-down priming, by itself, is subliminal (and in the brain unconscious), even though it can facilitate supraliminal processing of a subsequent expected event.

25. SEARCH, GENERALIZATION, AND NEUROBIOLOGICAL CORRELATES

The criterion of an acceptable 2/3 Rule match is defined by a parameter ρ called *vigilance* (Carpenter and Grossberg, 1987a, 1992). The vigilance parameter is computed in the orienting subsystem $\mathcal{A}$. Vigilance weighs how similar an input exemplar must be to a top-down prototype in order for resonance to occur. Resonance occurs if $\rho|\mathbf{I}| - |\mathbf{X}^*| \leq 0$. This inequality says that the F_1 attentional focus $\mathbf{X}^*$ inhibits $\mathcal{A}$ more than the input $\mathbf{I}$ excites it. If $\mathcal{A}$ remains quiet, then an $F_1 \leftrightarrow F_2$ resonance can develop.

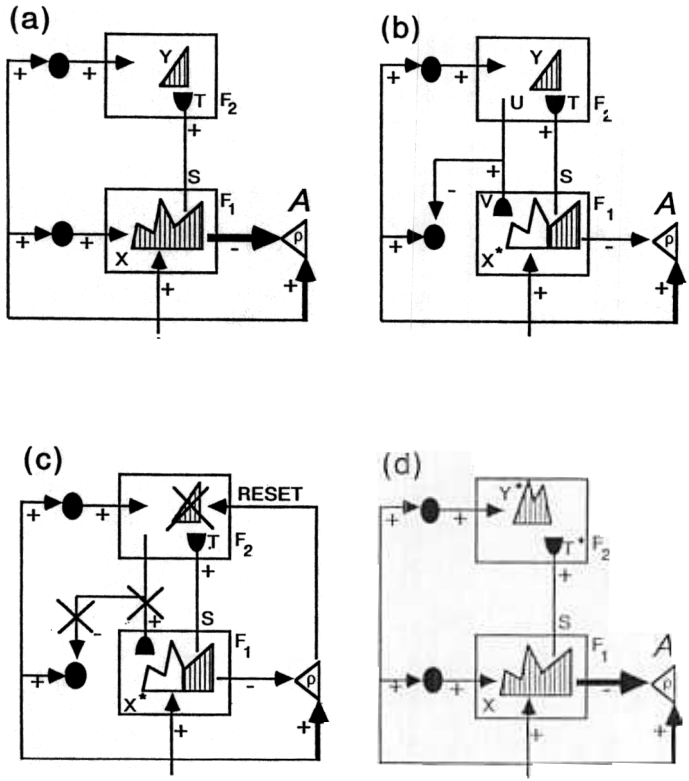


Figure 7. ART search for an F_2 recognition code: (a) The input pattern **I** generates the specific STM activity pattern **X** at F_1 as it nonspecifically activates the orienting subsystem $\mathcal{A}$. **X** is represented by the hatched pattern across F_1. Pattern **X** both inhibits $\mathcal{A}$ and generates the output pattern **S**. Pattern **S** is transformed by the LTM traces into the input pattern **T**, which activates the STM pattern **Y** across F_2. (b) Pattern **Y** generates the top-down output pattern **U** which is transformed into the prototype pattern **V**. If **V** mismatches **I** at F_1, then a new STM activity pattern $\mathbf{X}^*$ is generated at F_1. $\mathbf{X}^*$ is represented by the hatched pattern. Inactive nodes corresponding to **X** are unhatched. The reduction in total STM activity which occurs when **X** is transformed into $\mathbf{X}^*$ causes a decrease in the total inhibition from F_1 to $\mathcal{A}$. (c) If the vigilance criterion fails to be met, $\mathcal{A}$ releases a nonspecific arousal wave to F_2, which resets the STM pattern **Y** at F_2. (d) After **Y** is inhibited, its top-down prototype signal is eliminated, and **X** can be reinstated at F_1. Enduring traces of the prior reset lead **X** to activate a different STM pattern $\mathbf{Y}^*$ at F_2. If the top-down prototype due to $\mathbf{Y}^*$ also mismatches **I** at F_1, then the search for an appropriate F_2 code continues until a more appropriate F_2 representation is selected. Then an attentive resonance develops and learning of the attended data is initiated. (Reprinted with permission from Carpenter, Grossberg, and Rosen, 1991.)

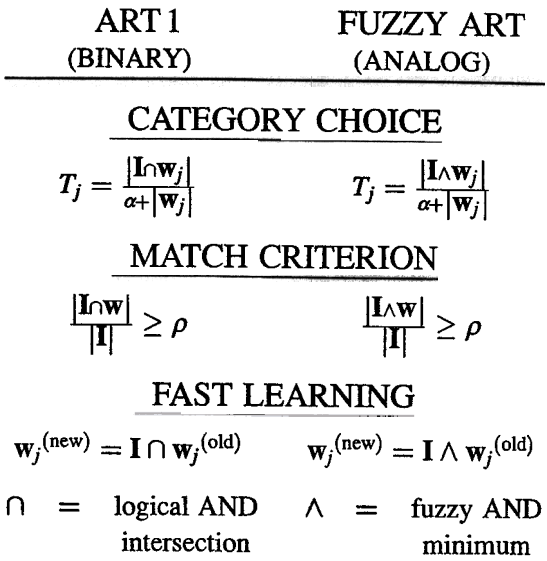


Figure 8. Comparison of ART 1 and Fuzzy ART. (Reprinted with permission from Carpenter, Grossberg, and Rosen, 1991.)

Vigilance calibrates how much novelty the system can tolerate before activating $\mathcal{A}$ and searching for a different category. If the top-down expectation and the bottom-up input are too different to satisfy the resonance criterion, then hypothesis testing, or memory search, is triggered. Memory search leads to selection of a better category at level F_2 with which to represent the input features at level F_1. During search, the orienting subsystem interacts with the attentional subsystem, as in Figures 7c and 7d, to rapidly reset mismatched categories and to select other F_2 representations with which to learn about novel events, without risking unselective forgetting of previous knowledge. Search may select a familiar category if its prototype is similar enough to the input to satisfy the vigilance criterion. The prototype may then be refined by 2/3 Rule attentional focussing. If the input is too different from any previously learned prototype, then an uncommitted population of F_2 cells is selected and learning of a new category is initiated.

Because vigilance can vary across learning trials, recognition categories capable of encoding widely differing degrees of generalization or abstraction can be learned by a single ART system. Low vigilance leads to broad generalization and abstract prototypes. High vigilance leads to narrow generalization and to prototypes that represent fewer input exemplars, even a single exemplar. Thus a single ART system may be used, say, to recognize abstract categories of faces and dogs, as well as individual faces and dogs. A single system can learn both, as the need arises, by increasing vigilance just enough to activate $\mathcal{A}$ if a previous categorization leads to a predictive error (Carpenter and Grossberg, 1992; Carpenter, Grossberg, and Reynolds, 1991; Carpenter, Grossberg, Markuzon, Reynolds, and Rosen, 1992). ART systems hereby provide a new answer to whether the brain learns

prototypes or exemplars. Various authors have realized that neither one nor the other alternative is satisfactory, and that a hybrid system is needed (Smith, 1990). ART systems can perform this hybrid function in a manner that is sensitive to environmental demands.

These properties of ART systems have been used to explain and predict a variety of cognitive and brain data that have, as yet, received no other theoretical explanation (Carpenter and Grossberg, 1991; Grossberg, 1987a, 1987b). For example, a formal lesion of the orienting subsystem creates a memory disturbance that remarkably mimics properties of medial temporal amnesia (Carpenter and Grossberg, 1987c, 1993; Grossberg and Merrill, 1992). These and related data correspondences to orienting properties (Grossberg and Merrill, 1992) have led to a neurobiological interpretation of the orienting subsystem in terms of the hippocampal formation of the brain. In applications to visual object recognition, the interactions within the F_1 and F_2 levels of the attentional subsystem are interpreted in terms of data concerning the prestriate visual cortex and the inferotemporal cortex (Desimone, 1992), with the attentional gain control pathway interpreted in terms of the pulvinar region of the brain. The ability of ART systems to form categories of variable generalization is linked to the ability of inferotemporal cortex to form both particular (exemplar) and general (prototype) visual representations.

26. A CONNECTION BETWEEN ART SYSTEMS AND FUZZY LOGIC

Fuzzy ART is a generalization of ART 1 that incorporates operations from fuzzy logic (Carpenter, Grossberg, and Rosen, 1991). Although ART 1 can learn to classify only binary input patterns, Fuzzy ART can learn to classify both analog and binary input patterns. Moreover, Fuzzy ART reduces to ART 1 in response to binary input patterns. As shown in Figure 8, the generalization to learning both analog and binary input patterns is achieved by replacing appearances of the intersection operator ($\cap$) in ART 1 by the MIN operator ($\wedge$) of fuzzy set theory. The MIN operator reduces to the intersection operator in the binary case. Of particular interest is the fact that, as parameter α approaches 0, the function T_j which controls category choice through the bottom-up filter reduces to the operation of fuzzy subsethood (Kosko, 1986). T_j then measures the degree to which the adaptive weight vector $\mathbf{w}_j$ is a fuzzy subset of the input vector $\mathbf{I}$.

In Fuzzy ART, input vectors are normalized at a preprocessing stage (Figure 9). This normalization procedure, called complement coding, leads to a symmetric theory in which the MIN operator ($\wedge$) and the MAX operator ($\vee$) of fuzzy set theory (Zadeh, 1965) play complementary roles. The categories formed by Fuzzy ART are then hyper-rectangles. Figure 10 illustrates how MIN and MAX define these rectangles in the 2-dimensional case. The MIN and MAX values define the acceptable range of feature variation in each dimension. Complement coding uses on-cells (with activity $\mathbf{a}$ in Figure 9) and off-cells (with activity $\mathbf{a}^c$ in Figure 9) to represent the input pattern, and preserves individual feature amplitudes while normalizing the total on-cell/off-cell vector. The on-cell portion of a prototype encodes features that are critically present in category exemplars, while the off-cell portion encodes features that are critically absent. Each category is then defined by an interval of expected values for each input feature. For instance, Fuzzy ART would encode the feature of "hair on head" by a wide interval ([A, 1]) for the category "man", whereas the feature "hat on head" would be encoded by a wide interval ([0, B]). On the other hand, the category "dog" would be encoded by two narrow intervals, [C, 1] for hair and [0, D] for hat, corresponding to narrower ranges of expectations for these two features.

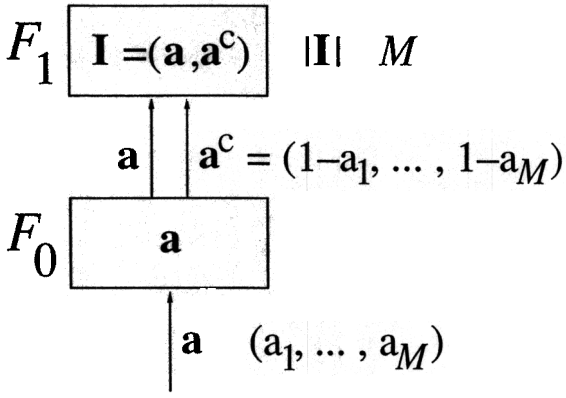
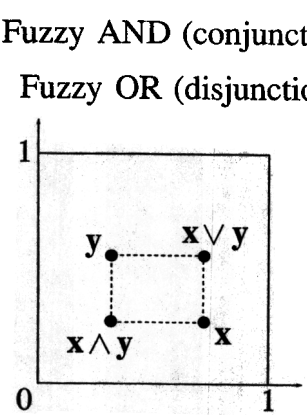


Figure 9. Complement coding uses on-cell and off-cell pairs to normalize input vectors. (Reprinted with permission from Carpenter, Grossberg, and Rosen, 1991.)

Learning in Fuzzy ART is stable because all adaptive weights can only decrease in time. Decreasing weights correspond to increasing sizes of category "boxes". This theorem is proved in Carpenter, Grossberg, and Rosen (1991). Smaller vigilance values lead to larger category boxes. Learning stops when the input space is covered by boxes. The use of complement coding works with the property of increasing box size to prevent a proliferation of categories. With fast learning, constant vigilance, and a finite input set of arbitrary size and composition, it has been proved that learning stabilizes after just one presentation of each input pattern. A fast-commit slow-recode option combines fast learning with a forgetting rule that buffers system memory against noise. Using this option, rare events can be rapidly learned, yet previously learned memories are not rapidly erased in response to statistically unreliable input fluctuations. The equations that define the Fuzzy ART algorithm are listed in Section 29.

27. FUZZY ARTMAP AND FUSION ARTMAP: SUPERVISED INCREMENTAL LEARNING, CATEGORIZATION, AND PREDICTION

Individual ART modules typically learn in an unsupervised mode. ART systems capable of supervised learning, categorization, and prediction have also recently been introduced (Asfour, Carpenter, Grossberg, and Lesher, 1993; Carpenter and Grossberg, 1992; Carpenter, Grossberg, and Reynolds, 1991; Carpenter, Grossberg, Markuzon, Reynolds, and Rosen, 1992; Carpenter, Grossberg, and Iizuka, 1992). Unlike many supervised learning networks, such as back propagation, these ART systems are capable of functioning in either an unsupervised or supervised mode, depending on whether environmental feedback is available. When supervised learning of Fuzzy ART controls category formation, a predictive error can force the creation of new categories that could not otherwise be learned due to monotone increase in category size through time in the unsupervised case. Supervision permits the creation of complex categorical structures without a loss of stability. The main additional ingredients whereby Fuzzy ART modules are combined into a supervised ART architectures are now summarized.



$$\mathbf{x} = (x_1, x_2) \qquad \mathbf{y} = (y_1, y_2)$$
$$(\mathbf{x} \wedge \mathbf{y})_1 = \min(x_1, y_1) \qquad (\mathbf{x} \wedge \mathbf{y})_2 = \min(x_2, y_2)$$
$$(\mathbf{x} \vee \mathbf{y})_1 = \max(x_1, y_1) \qquad (\mathbf{x} \vee \mathbf{y})_2 = \max(x_2, y_2)$$

Figure 10. Fuzzy AND and OR operations generate category hyper-rectangles. (Reprinted with permission from Carpenter, Grossberg, and Rosen, 1991.)

The simplest supervised ART systems are generically called ARTMAP. An ARTMAP that is built up from Fuzzy ART modules is called a Fuzzy ARTMAP system.

Each Fuzzy ARTMAP system includes a pair of Fuzzy ART modules (ART_a and ART_b), as in Figure 11. During supervised learning, ART_a receives a stream $\{\mathbf{a}^{(p)}\}$ of input patterns and ART_b receives a stream $\{\mathbf{b}^{(p)}\}$ of input patterns, where $\mathbf{b}^{(p)}$ is the correct prediction given $\mathbf{a}^{(p)}$. These modules are linked by an associative learning network and an internal controller that ensures autonomous system operation in real time. The controller is designed to create the minimal number of ART_a recognition categories, or "hidden units," needed to meet accuracy criteria. As noted above, this is accomplished by realizing a Minimax Learning Rule that conjointly minimizes predictive error and maximizes predictive generalization. This scheme automatically links predictive success to category size on a trial-by-trial basis using only local operations. It works by increasing the vigilance parameter ρ_a of ART_a by the minimal amount needed to correct a predictive error at ART_b (Figure 12).

Parameter ρ_a calibrates the minimum confidence that ART_a must have in a recognition category, or hypothesis, that is activated by an input $\mathbf{a}^{(p)}$ in order for ART_a to accept that category, rather than search for a better one through an automatically controlled process of hypothesis testing. As in ART 1, lower values of ρ_a enable larger categories to form. These lower ρ_a values lead to broader generalization and higher code compression. A predictive failure at ART_b increases the minimal confidence ρ_a by the least amount needed to trigger hypothesis testing at ART_a, using a mechanism called *match tracking* (Carpenter, Grossberg, and Reynolds, 1991). Match tracking sacrifices the minimum amount of generalization necessary to correct the predictive error. Speaking intuitively,

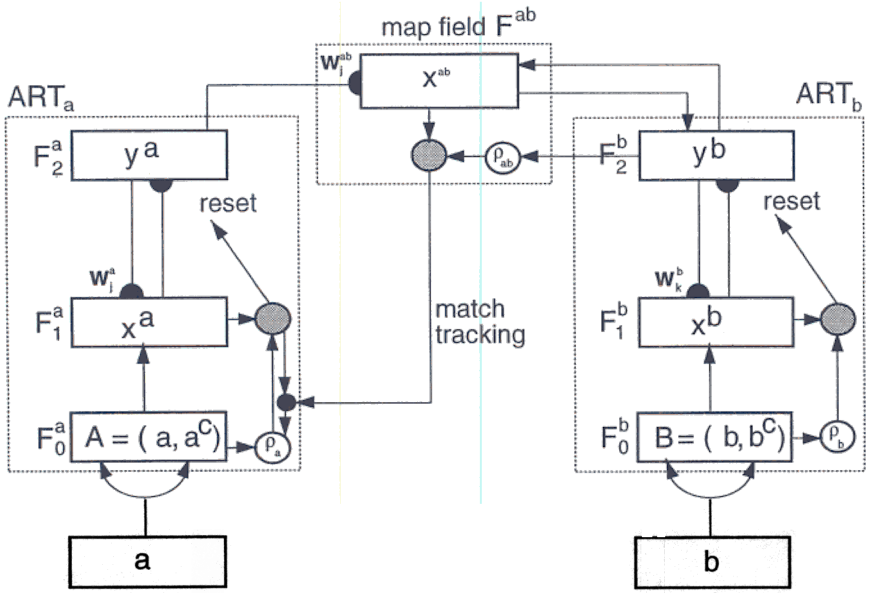


Figure 11. Fuzzy ARTMAP architecture. The ART_a complement coding preprocessor transforms the M_a-vector $\mathbf{a}$ into the $2M_a$-vector $\mathbf{A} = (\mathbf{a}, \mathbf{a}^c)$ at the ART_a field F_0^a. $\mathbf{A}$ is the input vector to the ART_a field F_1^a. Similarly, the input to F_1^b is the $2M_b$-vector $(\mathbf{b}, \mathbf{b}^c)$. When a prediction by ART_a is disconfirmed at ART_b, inhibition of map field activation induces the match tracking process. Match tracking raises the ART_a vigilance ρ_a to just above the F_1^a to F_0^a match ratio $|\mathbf{x}^a|/|\mathbf{A}|$. This triggers an ART_a search which leads to activation of either an ART_a category that correctly predicts $\mathbf{b}$ or to a previously uncommitted ART_a category node. (Reprinted with permission from Carpenter, Grossberg, Markuzon, Reynolds, and Rosen, 1992.)

match tracking operationalizes the idea that the system must have accepted hypotheses with too little confidence to satisfy the demands of a particular environment. Match tracking increases the criterion confidence just enough to trigger hypothesis testing. Hypothesis testing leads to the selection of a new ART_a category, which focuses attention on a new cluster of $\mathbf{a}^{(p)}$ input features that is better able to predict $\mathbf{b}^{(p)}$. Due to the combination of match tracking and fast learning, a single ARTMAP system can learn a different prediction for a rare event than for a cloud of similar frequent events in which it is embedded.

A generalization of Fuzzy ARTMAP, called Fusion ARTMAP, has also recently been introduced to handle multidimensional data fusion, classification, and prediction problems (Asfour, Carpenter, Grossberg, and Lesher, 1993). In Fusion ARTMAP, multiple data channels process different sorts of input vectors in their own ART modules before all

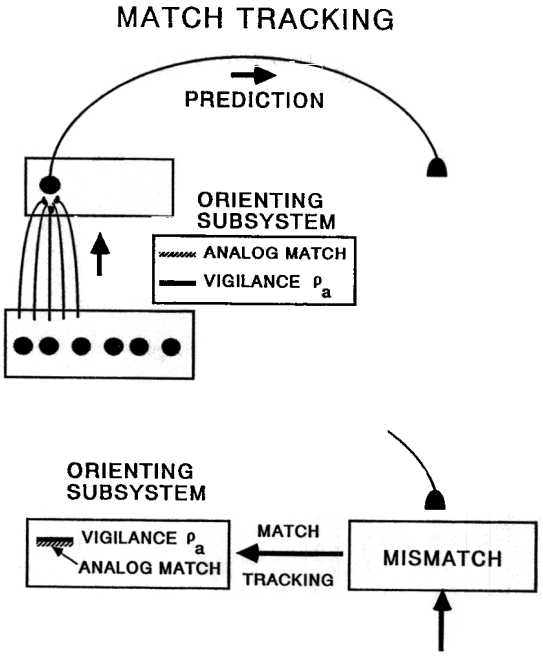


Figure 12. Match tracking: (a) A prediction is made by ART_a when the baseline vigilance ρ_a is less than the analog match value. (b) A predictive error at ART_b increases the baseline vigilance value of ART_a until it just exceeds the analog match value, and thereby triggers hypothesis testing that searches for a more predictive bundle of features to which to attend. (Reprinted with permission from Carpenter and Grossberg, 1992.)

the ART modules cooperate to form a global classification and prediction. A predictive error simultaneously raises the vigilance parameters of all the component ART modules. The module with the poorest match of input to prototype is driven first to reset and search. As a result, the channels whose data are classified with the least confidence are searched before more confident classifications are reset. Channels which provide good data matches may thus not need to create new categories just because other channels exhibit poor matches. Using this parallel match tracking scheme, the network selectively improves learning where it is poor, while sparing the learning that is good. Such an automatic credit assignment has been shown in benchmark studies to generate more persimonious classifications of multidimensional data than are learned by a one-channel Fuzzy ARTMAP. Two benchmark studies using Fuzzy ARTMAP are summarized below to show that even a one-channel network has powerful classification capabilities.

28. TWO BENCHMARK STUDIES: LETTER AND WRITTEN DIGIT RECOGNITION

As summarized in Table 1, Fuzzy ARTMAP has been benchmarked against a variety of machine learning, neural network, and genetic algorithms with considerable success.

ARTMAP BENCHMARK STUDIES

1. Medical database - mortality following coronary bypass grafting (CABG) surgery
   FUZZY ARTMAP significantly outperforms
   LOGISTIC REGRESSION
   ADDITIVE MODEL
   BAYESIAN ASSIGNMENT
   CLUSTER ANALYSIS
   CLASSIFICATION AND REGRESSION TREES
   EXPERT PANEL-DERIVED SICKNESS SCORES
   PRINCIPAL COMPONENT ANALYSIS
2. Mushroom database
   DECISION TREES (90-95% correct)
   ARTMAP (100% correct)
   Training set an order of magnitude smaller
3. Letter recognition database
   GENETIC ALGORITHM (82% correct)
   FUZZY ARTMAP (96% correct)
4. Circle-in-the-Square task
   BACK PROPAGATION (90% correct)
   FUZZY ARTMAP (99.5% correct)
5. Two-Spiral task
   BACK PROPAGATION (10,000-20,000 training epochs)
   FUZZY ARTMAP (1-5 training epochs)

Table 1

An illustrative study used a benchmark machine learning task that Frey and Slate (1991) developed and described as a "difficult categorization problem" (p. 161). The task requires a system to identify an input exemplar as one of 26 capital letters A–Z. The database was derived from 20,000 unique black-and-white pixel images. The difficulty of the task is due to the wide variety of letter types represented: the twenty "fonts represent five different stroke styles (simplex, duplex, complex, and Gothic) and six different letter styles (block, script, italic, English, Italian, and German)" (p. 162). In addition each image was randomly distorted, leaving many of the characters misshapen. Sixteen numerical feature attributes were then obtained from each character image, and each attribute value was scaled to a range of 0 to 15. The resulting Letter Image Recognition file is archived in the UCI Repository of Machine Learning Databases and Domain Theories, maintained by David Aha and Patrick Murphy (ml_repository@ics.uci.edu).

Frey and Slate used this database to test performance of a family of classifiers based on Holland's genetic algorithms (Holland, 1980). The training set consisted of 16,000 exemplars, with the remaining 4,000 exemplars used for testing. Genetic algorithm classifiers having different input representations, weight update and rule creation schemes, and system parameters were systematically compared. Training was carried out for 5 epochs, plus a sixth "verification" pass during which no new rules were created but a large number

of unsatisfactory rules were discarded. In Frey and Slate's comparative study, these systems had correct prediction rates that ranged from 24.5% to 80.8% on the 4,000-item test set. The best performance (80.8%) was obtained using an integer input representation, a reward sharing weight update, an exemplar method of rule creation, and a parameter setting that allowed an unused or erroneous rule to stay in the system for a long time before being discarded. After training, the optimal case, that had 80.8% performance rate, ended with 1,302 rules and 8 attributes per rule, plus over 35,000 more rules that were discarded during verification. (For purposes of comparison, a rule is somewhat analogous to an ART_a category in ARTMAP, and the number of attributes per rule is analogous to the size $|\mathbf{w}_j^a|$ of ART_a category weight vectors.) Building on the results of their comparative study, Frey and Slate investigated two types of alternative algorithms, namely an accuracy-utility bidding system, that had slightly improved performance (81.6%) in the best case; and an exemplar/hybrid rule creation scheme that further improved performance, to a maximum of 82.7%, but that required the creation of over 100,000 rules prior to the verification step.

Fuzzy ARTMAP had an error rate on the letter recognition task that was consistently less than one third that of the three best Frey-Slate genetic algorithm classifiers described above. In particular, after 1 to 5 epochs, individual Fuzzy ARTMAP systems had a robust prediction rate of 90% to 94% on the 4,000-item test set. A *voting strategy* consistently improved this performance. This voting strategy is based on the observation that ARTMAP fast learning typically leads to different adaptive weights and recognition categories for different orderings of a given training set, even when overall predictive accuracy of all simulations is similar. The different category structures cause the set of test items where errors occur to vary from one simulation to the next. The voting strategy uses an ARTMAP system that is trained several times on input sets with different orderings. The final prediction for a given test set item is the one made by the largest number of simulations. Since the set of items making erroneous predictions varies from one simulation to the next, voting cancels many of the errors. Such a voting strategy can also be used to assign confidence estimates to competing predictions given small, noisy, or incomplete training sets. Voting consistently eliminated 25%–43% of the errors, giving a robust prediction rate of 92%–96%. Moreover Fuzzy ARTMAP simulations each created fewer than 1,070 ART_a categories, compared to the 1,040–1,302 final rules of the three genetic classifiers with the best performance rates. Most Fuzzy ARTMAP learning occurred on the first epoch, with test set performance on systems trained for one epoch typically over 97% that of systems exposed to inputs for five epochs.

Rapid learning was also found in a benchmark study of written digit recognition, where the correct prediction rate on the test set after one epoch reached over 99% of its best performance (Carpenter, Grossberg, and Iizuka, 1992). In this study, Fuzzy ARTMAP was tested along with back propagation and a self-organizing feature map. Voting yielded Fuzzy ARTMAP average performance rates on the test set of 97.4% after an average number of 4.6 training epochs. Back propagation achieved its best average performance rates of 96% after 100 training epochs. Self-organizing feature maps achieved a best level of 96.5%, again after many training epochs.

In summary, on a variety of benchmarks (see also Table 1, Carpenter, Grossberg, and Reynolds, 1991, and Carpenter *et al.*, 1992), Fuzzy ARTMAP has demonstrated either much faster learning, better performance, or both, than alternative machine learning,

ARTMAP

ARTMAP can autonomously learn about

(A) RARE EVENTS
Need FAST learning

(B) LARGE NONSTATIONARY DATABASES
Need STABLE learning

(C) MORPHOLOGICALLY VARIABLE EVENTS
Need MULTIPLE SCALES of generalization (fine/coarse)

(D) ONE-TO-MANY AND MANY-TO-ONE RELATIONSHIPS
Need categorization, naming, and expert knowledge

To realize these properties ARTMAP systems:

(E) PAY ATTENTION
Ignore masses of irrelevant data

(F) TEST HYPOTHESES
Discover predictive constraints hidden in data streams

(G) CHOOSE BEST ANSWERS
Quickly select globally optimal solution at any stage of learning

(H) CALIBRATE CONFIDENCE
Measure on-line how well a hypothesis matches the data

(I) DISCOVER RULES
Identify transparent IF-THEN relations at each learning stage

(J) SCALE
Preserve all desirable properties in arbitrarily large problems

Table 2

genetic, or neural network algorithms. Perhaps more importantly, Fuzzy ARTMAP can be used in an important class of applications where many other adaptive pattern recognition algorithms cannot perform well (see Table 2). These are the applications where very large nonstationary databases need to be rapidly organized into stable variable-compression categories under real-time autonomous learning conditions.

29. SUMMARY OF THE FUZZY ART ALGORITHM

ART field activity vectors: Each ART system includes a field F_0 of nodes that represent a current input vector; a field F_1 that receives both bottom-up input from F_0 and top-down input from a field F_2 that represents the active code, or category. The F_0 activity vector is denoted $\mathbf{I} = (I_1, \ldots, I_M)$, with each component I_i in the interval [0,1], $i = 1, \ldots, M$. The F_1 activity vector is denoted $\mathbf{x} = (x_1, \ldots, x_M)$ and the F_2 activity vector is denoted $\mathbf{y} = (y_1, \ldots, y_N)$. The number of nodes in each field is arbitrary.

Weight vector: Associated with each F_2 category node $j (j = 1, \ldots, N)$ is a vector

$\mathbf{w}_j \equiv (w_{j1}, \ldots, w_{jM})$ of adaptive weights, or LTM traces. Initially

$$w_{j1}(0) = \ldots = w_{jM}(0) = 1; \tag{96}$$

then each category is said to be *uncommitted.* After a category is selected for coding it becomes *committed.* As shown below, each LTM trace w_{ji} is monotone nonincreasing through time and hence converges to a limit. The Fuzzy ART weight vector $\mathbf{w}_j$ subsumes both the bottom-up and top-down weight vectors of ART 1.

Parameters: Fuzzy ART dynamics are determined by a choice parameter $\alpha > 0$; a learning rate parameter $\beta \in [0,1]$; and a vigilance parameter $\rho \in [0,1]$.

Category choice: For each input $\mathbf{I}$ and F_2 node j, the *choice function* T_j is defined by

$$T_j(\mathbf{I}) = \frac{|\mathbf{I} \wedge \mathbf{w}_j|}{\alpha + |\mathbf{w}_j|}, \tag{97}$$

where the fuzzy AND operator $\wedge$ is defined by

$$(\mathbf{p} \wedge \mathbf{q})_i \equiv \min(p_i, q_i) \tag{98}$$

and where the norm $|\cdot|$ is defined by

$$|\mathbf{p}| \equiv \sum_{i=1}^{M} |p_i|. \tag{99}$$

for any M-dimensional vectors $\mathbf{p}$ and $\mathbf{q}$. For notational simplicity, $T_j(\mathbf{I})$ in (97) is often written as T_j when the input $\mathbf{I}$ is fixed.

The system is said to make a *category choice* when at most one F_2 node can become active at a given time. The category choice is indexed by J, where

$$T_J = \max\{T_j : j = 1 \ldots N\}. \tag{100}$$

If more than one T_j is maximal, the category j with the smallest index is chosen. In particular, nodes become committed in order $j = 1, 2, 3, \ldots$. When the J^{th} category is chosen, $y_J = 1$; and $y_j = 0$ for $j \neq J$. In a choice system, the F_1 activity vector $\mathbf{x}$ obeys the equation

$$\mathbf{x} = \begin{cases} \mathbf{I} & \text{if } F_2 \text{ is inactive} \\ \mathbf{I} \wedge \mathbf{w}_J & \text{if the } J^{th} F_2 \text{ node is chosen.} \end{cases} \tag{101}$$

Resonance or reset: *Resonance* occurs if the *match function* $|\mathbf{I} \wedge \mathbf{w}_J|/|\mathbf{I}|$ of the chosen category meets the vigilance criterion:

$$\frac{|\mathbf{I} \wedge \mathbf{w}_J|}{|\mathbf{I}|} \geq \rho;$$

that is, by (6), when the J^{th} category is chosen, resonance occurs if

$$|\mathbf{x}| = |\mathbf{I} \wedge \mathbf{w}_J| \geq \rho |\mathbf{I}|.$$

Learning then ensues, as defined below. *Mismatch reset* occurs if

$$\frac{|\mathbf{I} \wedge \mathbf{w}_J|}{|\mathbf{I}|} < \rho; \tag{104}$$

that is, if

$$|\mathbf{x}| = |\mathbf{I} \wedge \mathbf{w}_J| < \rho|\mathbf{I}|. \tag{105}$$

Then the value of the choice function T_J is set to 0 for the duration of the input presentation to prevent the persistent selection of the same category during search. A new index J is then chosen, by (100). The search process continues until the chosen J satisfies (102).

Learning: Once search ends, the weight vector $\mathbf{w}_J$ is updated according to the equation

$$\mathbf{w}_J^{(\text{new})} = \beta(\mathbf{I} \wedge \mathbf{w}_J^{(\text{old})}) + (1 - \beta)\mathbf{w}_J^{(\text{old})}. \tag{106}$$

Fast learning corresponds to setting $\beta = 1$. The learning law used in the EACH system of Salzberg (1990) is equivalent to equation (106) in the fast-learn limit with the complement coding option described below.

Fast-commit slow-recode option: For efficient coding of noisy input sets, it is useful to set $\beta = 1$ when J is an uncommitted node, and then to take $\beta < 1$ after the category is committed. Then $\mathbf{w}_J^{(\text{new})} = \mathbf{I}$ the first time category J becomes active. Moore (1989) introduced the learning law (106), with fast commitment and slow recoding, to investigate a variety of generalized ART 1 models. Some of these models are similar to Fuzzy ART, but none includes the complement coding option. Moore described a category proliferation problem that can occur in some analog ART systems when a large number of inputs erode the norm of weight vectors. Complement coding solves this problem.

Input normalization/complement coding option: Proliferation of categories is avoided in Fuzzy ART if inputs are normalized. *Complement coding* is a normalization rule that preserves amplitude information. Complement coding represents both the on-response and the off-response to an input vector $\mathbf{a}$ (Figure 8). To define this operation in its simplest form, let $\mathbf{a}$ itself represent the on-response. The complement of $\mathbf{a}$, denoted by $\mathbf{a}^c$, represents the off-response, where

$$a_i^c \equiv 1 - a_i. \tag{107}$$

The complement coded input $\mathbf{I}$ to the field F_1 is the 2M-dimensional vector

$$\mathbf{I} = (\mathbf{a}, \mathbf{a}^c) \equiv (a_1, \ldots, a_M, a_1^c, \ldots, a_M^c). \tag{108}$$

Note that

$$\begin{aligned} |\mathbf{I}| &= |(\mathbf{a}, \mathbf{a}^c)| \\ &= \sum_{i=1}^{M} a_i + (M - \sum_{i=1}^{M} a_i) \\ &= M, \end{aligned} \tag{109}$$

so inputs preprocessed into complement coding form are automatically normalized. Where complement coding is used, the initial condition (96) is replaced by

$$w_{j1}(0) = \ldots = w_{j,2M}(0) = 1. \tag{110}$$

30. FUZZY ARTMAP ALGORITHM

The Fuzzy ARTMAP system incorporates two Fuzzy ART modules ART_a and ART_b that are linked together via an inter-ART module F^{ab} called a *map field.* The map field is used to form predictive associations between categories and to realize the *match tracking rule* whereby the vigilance parameter of ART_a increases in response to a predictive mismatch at ART_b. The interactions mediated by the map field F^{ab} may be operationally characterized as follows.

ART_a and ART_b

Inputs to ART_a and ART_b are in the complement code form: for ART_a, $\mathbf{I} = \mathbf{A} = (\mathbf{a}, \mathbf{a}^c)$; for ART_b, $\mathbf{I} = \mathbf{B} = (\mathbf{b}, \mathbf{b}^c)$ (Figure 10). Variables in ART_a or ART_b are designated by subscripts or superscripts "a" or "b". For ART_a, let $\mathbf{x}^a \equiv (x_1^a \ldots x_{2M_a}^a)$ denote the F_1^a output vector; let $\mathbf{y}^a \equiv (y_1^a \ldots y_{N_a}^a)$ denote the F_2^a output vector; and let $\mathbf{w}_j^a \equiv (w_{j1}^a, w_{j2}^a, \ldots, w_{j,2M_a}^a)$ denote the j^{th} ART_a weight vector. For ART_b, let $\mathbf{x}^b \equiv (x_1^b \ldots x_{2M_b}^b)$ denote the F_1^b output vector; let $\mathbf{y}^b \equiv (y_1^b \ldots y_{N_b}^b)$ denote the F_2^b output vector; and let $\mathbf{w}_k^b \equiv (w_{k1}^b, w_{k2}^b, \ldots, w_{k,2M_b}^b)$ denote the k^{th} ART_b weight vector. For the map field, let $\mathbf{x}^{ab} \equiv (x_1^{ab}, \ldots, x_{N_b}^{ab})$ denote the F^{ab} output vector, and let $\mathbf{w}_j^{ab} \equiv (w_{j1}^{ab}, \ldots, w_{jN_b}^{ab})$ denote the weight vector from the j^{th} F_2^a node to F^{ab}. Vectors $\mathbf{x}^a, \mathbf{y}^a, \mathbf{x}^b, \mathbf{y}^b$, and $\mathbf{x}^{ab}$ are set to $\mathbf{0}$ between input presentations.

Map field activation

The map field F^{ab} is activated whenever one of the ART_a or ART_b categories is active. If node J of F_2^a is chosen, then its weights $\mathbf{w}_J^{ab}$ activate F^{ab}. If node K in F_2^b is active, then the node K in F^{ab} is activated by 1–to–1 pathways between F_2^b and F^{ab}. If both ART_a and ART_b are active, then F^{ab} becomes active only if ART_a predicts the same category as ART_b via the weights $\mathbf{w}_j^{ab}$. The F^{ab} output vector $\mathbf{x}^{ab}$ obeys

$$\mathbf{x}^{ab} = \begin{cases} \mathbf{y}^b \wedge \mathbf{w}_J^{ab} & \text{if the Jth } F_2^a \text{ node is active and } F_2^b \text{ is active} \\ \mathbf{w}_J^{ab} & \text{if the Jth } F_2^a \text{ node is active and } F_2^b \text{ is inactive} \\ \mathbf{y}^b & \text{if } F_2^a \text{ is inactive and } F_2^b \text{ is active} \\ \mathbf{0} & \text{if } F_2^a \text{ is inactive and } F_2^b \text{ is inactive.} \end{cases} \tag{111}$$

By (111), $\mathbf{x}^{ab} = \mathbf{0}$ if the prediction $\mathbf{w}_J^{ab}$ is disconfirmed by $\mathbf{y}^b$. Such a mismatch event triggers an ART_a search for a better category, as follows.

Match tracking

At the start of each input presentation the ART_a vigilance parameter ρ_a equals a baseline vigilance $\overline{\rho_a}$. The map field vigilance parameter is ρ_{ab}. If

$$|\mathbf{x}^{ab}| < \rho_{ab}|\mathbf{y}^b|, \tag{112}$$

then ρ_a is increased until it is slightly larger than $|\mathbf{A} \wedge \mathbf{w}_J^a||\mathbf{A}|^{-1}$, where $\mathbf{A}$ is the input to F_1^a, in complement coding form. Then

$$|\mathbf{x}^a| = |\mathbf{A} \wedge \mathbf{w}_J^a| < \rho_a|\mathbf{A}|, \tag{113}$$

where J is the index of the active F_2^a node, as in (105). When this occurs, ART_a search leads either to activation of another F_2^a node J with

$$|\mathbf{x}^a| = |\mathbf{A} \wedge \mathbf{w}_J^a| \geq \rho_a|\mathbf{A}| \tag{114}$$

and

$$|\mathbf{x}^{ab}| = |\mathbf{y}^b \wedge \mathbf{w}_J^{ab}| \geq \rho_{ab}|\mathbf{y}^b|; \tag{115}$$

or, if no such node exists, to the shut-down of F_2^a for the remainder of the input presentation.

Map field learning

Learning rules determine how the map field weights w_{jk}^{ab} change through time, as follows. Weights w_{jk}^{ab} in $F_2^a \rightarrow F^{ab}$ paths initially satisfy

$$w_{jk}^{ab}(0) = 1.$$

During resonance with the ART_a category J active, $\mathbf{w}_J^{ab}$ approaches the map field vector $\mathbf{x}^{ab}$. With fast learning, once J learns to predict the ART_b category K, that association is permanent; i.e., $w_{JK}^{ab} = 1$ for all time.

REFERENCES

Adams, J.A. (1967). **Human memory**. New York: McGraw-Hill.

Amari, S.-I. and Arbib, M. (Eds.) (1982). **Competition and cooperation in neural networks**. New York, NY: Springer-Verlag.

Amari, S.-I. and Takeuchi, A. (1978). Mathematical theory on formation of category detecting nerve cells. *Biological Cybernetics*, **29**, 127–136.

Asch, S.E. and Ebenholtz, S.M. (1962). The principle of associative symmetry. *Proceedings of the American Philosophical Society*, **106**, 135–163.

Asfour, Y.R., Carpenter, G.A., Grossberg, S., and Lesher, G. (1993). Fusion ARTMAP: A neural network architecture for multi-channel data fusion and classification. Technical Report CAS/CNS TR93-004, Boston, MA: Boston University. Submitted for publication.

Bienenstock, E.L., Cooper, L.N., and Munro, P.W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, **2**, 32–48.

Bradski, G., Carpenter, G.A., and Grossberg, S. (1992). Working memory networks for learning multiple groupings of temporal order with application to 3-D visual object recognition. *Neural Computation*, **4**, 270–286.

Carpenter, G.A. and Grossberg, S. (1987a). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, **37**, 54–115.

Carpenter, G.A. and Grossberg, S. (1987b). ART 2: Stable self-organization of pattern recognition codes for analog input patterns. *Applied Optics*, **26**, 4919–4930.

Carpenter, G.A. and Grossberg, S. (1987c). Neural dynamics of category learning and recognition: Attention, memory consolidation, and amnesia. In S. Grossberg (Ed.), **The adaptive brain, I: Cognition, learning, reinforcement, and rhythm**. Amsterdam: Elsevier/North Holland, pp. 238–286.

Carpenter, G.A. and Grossberg, S. (Eds.) (1991). **Pattern recognition by self-organizing neural networks**. Cambridge, MA: MIT Press.

Carpenter, G.A. and Grossberg, S. (1992). Fuzzy ARTMAP: Supervised learning, recognition, and prediction by a self-organizing neural network. *IEEE Communications Magazine*, **30**, 38–49.

Carpenter, G.A. and Grossberg, S. (1993). Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions. Technical Report CAS/CNS TR-92-021. Boston, MA: Boston University. *Trends in Neurosciences*, in press.

Carpenter, G.A., Grossberg, S., Markuzon, M., Reynolds, J.H., and Rosen, D.B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Network*, **3**, 698–713.

Carpenter, G.A., Grossberg, S., and Reynolds, J.H. (1991). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, **4**, 565–588.

Carpenter, G.A., Grossberg, S., and Rosen, D.B. (1991). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, **4**, 759–771.

Carpenter, G.A., Grossberg, S., and Iizuka, K. (1992). Comparative performance measures of Fuzzy ARTMAP, learned vector quantization, and back propagation for handwritten character recognition. **Proceedings of the international joint conference on neural networks, I**, 794–799. Piscataway, NJ: IEEE Service Center.

Cohen, M.A. (1988). Sustained oscillations in a symmetric cooperative-competitive neural network: Disproof of a conjecture about a content addressable memory. *Neural Networks*, **1**, 217–221.

Cohen, M.A. (1990). The stability of sustained oscillations in symmetric cooperative-competitive networks. *Neural Networks*, **3**, 609–612.

Cohen, M.A. (1992). The construction of arbitrary stable dynamics in nonlinear neural networks. *Neural Networks*, **5**, 83–103.

Cohen, M.A. and Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC–13**, 815–826.

Cohen, M.A. and Grossberg, S. (1986). Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short term memory. *Human Neurobiology*, **5**, 1–22.

Cohen, M.A., Grossberg, S., and Pribe, C. (1993). A neural pattern generator that exhibits frequency-dependent bi-manual coordination effects and quadruped gait transitions. Technical Report CAS/CNS TR-93-004. Boston, MA: Boston University. Submitted for publication.

Cole, K.S. (1968). **Membranes, ions, and impulses**. Berkeley, CA: University of California Press.

Collins, A.M. and Loftus, E.F. (1975). A spreading-activation theory of semantic memory. *Psychological Review*, **82**, 407–428.

Commons, M.L., Grossberg, S., and Staddon, J.E.R. (Eds.) (1991). **Neural network models of conditioning and action**. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cornsweet, T.N. (1970). **Visual perception**. New York, NY: Academic Press.

Crick, F. and Koch, C. (1990). Some reflections on visual awareness. **Cold Spring Harbor symposium on quantitative biology, LV, The brain**, Plainview, NY: Cold Spring Harbor Laboratory Press, 953–962.

Desimone, R. (1992). Neural circuits for visual attention in the primate brain. In G.A. Carpenter and S. Grossberg (Eds.), **Neural networks for vision and image processing**. Cambridge, MA: MIT Press, pp. 343–364.

Dixon, T.R. and Horton, D.L. (1968). **Verbal behavior and general behavior theory**. Englewood Cliffs, NJ: Prentice-Hall.

Eckhorn, R. Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., and Reitbock, H.J. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex? *Biological Cybernetics*, 1988, **60**, 121–130.

Eckhorn, R. and Schanze, T. (1991). Possible neural mechanisms of feature linking in the visual system: Stimulus-locked and stimulus-induced synchronizations. In A.

Babloyantz (Ed.), **Self-organization, emerging properties, and learning**. New York, NY: Plenum Press, pp. 63–80.

Ellias, S. and Grossberg, S. (1975). Pattern formation, contrast control, and oscillations in the short term memory of shunting on-center off-surround networks. *Biological Cybernetics*, **20**, 69–98.

Frey, P.W. and Slate, D.J. (1991). Letter recognition using Holland-style adaptive classifiers. *Machine Learning*, **6**, 161–182.

Gaudiano, P. (1992a). A unified neural model of spatio-temporal processing in X and Y retinal ganglion cells. *Biological Cybernetics*, **67**, 11–21.

Gaudiano, P. (1992b). Toward a unified theory of spatio-temporal processing in the retina. In G. Carpenter and S. Grossberg, (Eds.). **Neural networks for vision and image processing**. Cambridge, MA: MIT Press, pp. 195–220.

Gaudiano, P. and Grossberg, S. (1991). Vector associative maps: Unsupervised real-time error-based learning and control of movement trajectories. *Neural Networks*, **4**, 147–183.

Geman, S. (1981). The law of large numbers in neural modelling. In S. Grossberg (Ed.), **Mathematical psychology and psychophysiology**. Providence, RI: American Mathematical Society, pp. 91–106.

Gottlieb, G. (Ed.) (1976). **Neural and behavioral specificity** (Vol. 3). New York, NY: Academic Press.

Gray, C.M., Konig, P., Engel, A.K., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, **338**, 334–337.

Gray, C.M. and Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences*, **86**, 1698–1702.

Grossberg, S. (1961). Senior Fellowship thesis, Dartmouth College.

Grossberg, S. (1964). **The theory of embedding fields with applications to psychology and neurophysiology**. New York: Rockefeller Institute for Medical Research.

Grossberg, S. (1967). Nonlinear difference-differential equations in prediction and learning theory. *Proceedings of the National Academy of Sciences*, **58**, 1329–1334.

Grossberg, S. (1968a). Some physiological and biochemical consequences of psychological postulates. *Proceedings of the National Academy of Sciences*, **60**, 758–765.

Grossberg, S. (1968b). Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity. *Proceedings of the National Academy of Sciences*, **59**, 368–372.

Grossberg, S. (1969a). Embedding fields: A theory of learning with physiological implications. *Journal of Mathematical Psychology*, **6**, 209–239.

Grossberg, S. (1969b). On learning, information, lateral inhibition, and transmitters. *Mathematical Biosciences*, **4**, 255–310.

Grossberg, S., (1969c). On the serial learning of lists. *Mathematical Biosciences*, **4**, 201–253.

Grossberg, S. (1969d). On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks. *Journal of Statistical Physics*, **1**, 319–350.

Grossberg, S. (1969e). Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, I. *Journal of Mathematics and Mechanics*, **19**, 53–91.

Grossberg, S. (1969f). On the production and release of chemical transmitters and related topics in cellular control. *Journal of Theoretical Biology*, **22**, 325–364.

Grossberg, S. (1969g) On variational systems of some nonlinear difference-differential equations. *Journal of Differential Equations*, **6**, 544–577.

Grossberg, S. (1970a). Neural pattern discrimination. *Journal of Theoretical Biology*, **27**, 291–337.

Grossberg, S. (1970b). Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, II. *Studies in Applied Mathematics*, **49**, 135–166.

Grossberg, S. (1971a). Pavlovian pattern learning by nonlinear neural networks. *Proceedings of the National Academy of Sciences*, **68**, 828–831.

Grossberg, S. (1971b). On the dynamics of operant conditioning. *Journal of Theoretical Biology*, **33**, 225–255.

Grossberg, S. (1972a). Pattern learning by functional-differential neural networks with arbitrary path weights. In K. Schmitt (Ed.), **Delay and functional-differential equations and their applications**. New York: Academic Press. Reprinted in S. Grossberg (1982), **Studies of mind and brain**, pp. 157–193, Boston, MA: Reidel Press.

Grossberg, S. (1972b). Neural expectation: Cerebellar and retinal analogs of cells fired by learnable or unlearned pattern classes. *Kybernetik*, **10**, 49–57.

Grossberg, S. (1972c). A neural theory of punishment and avoidance, I: Qualitative theory. *Mathematical Biosciences*, **15**, 39–67.

Grossberg, S. (1972d). A neural theory of punishment and avoidance, II: Quantitative theory. *Mathematical Biosciences*, **15**, 253–285.

Grossberg, S. (1973). Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, **52**, 217–257. Reprinted in S. Grossberg (1982), **Studies of mind and brain**, pp. 332–378, Boston, MA: Reidel Press.

Grossberg, S. (1974). Classical and instrumental learning by neural networks. In R. Rosen and F. Snell (Eds.), **Progress in theoretical biology**. New York: Academic Press. Reprinted in S. Grossberg (1982), **Studies of mind and brain**, pp. 65–156, Boston, MA: Reidel Press.

Grossberg, S. (1975). A neural model of attention, reinforcement, and discrimination learning. *International Review of Neurobiology*, 1975, **18**, 263–327. Reprinted in S. Grossberg (1982), **Studies of mind and brain**, pp. 229–295, Boston, MA: Reidel Press.

Grossberg, S. (1976a). On the development of feature detectors in the visual cortex with applications to learning and reaction-diffusion systems. *Biological Cybernetics*, **21**, 145–159.

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, **23**, 121–

134.

Grossberg, S. (1976c). Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, **23**, 187–202.

Grossberg, S. (1976d). On the Development of feature detectors in the visual cortex with applications to learning and reaction-diffusion systems. *Biological Cybernetics*, **21**, 145–159.

Grossberg, S. (1978a). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen and F. Snell (Eds.), **Progress in theoretical biology**, Vol. 5. New York: Academic Press. Reprinted in S. Grossberg (1982), **Studies of mind and brain**, pp. 498-639, Boston, MA: Reidel Press.

Grossberg, S. (1978b). Behavioral contrast in short term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology*, **3**, 199–219.

Grossberg, S. (1978c). Decisions, patterns, and oscillations in nonlinear competitive systems with applications to Volterra-Lotka systems. *Journal of Theoretical Biology*, **73**, 101–130.

Grossberg, S. (1978d). Competition, decision, and consensus. *Journal of Mathematical Analysis and Applications*, **66**, 470–493.

Grossberg, S. (1980a). How does a brain build a cognitive code? *Psychological Review*, **1**, 1–51.

Grossberg, S. (1980b). Intracellular mechanisms of adaptation and self-regulation in self-organizing networks: The role of chemical transducers. *Bulletin of Mathematical Biology*, **42**, 365–396.

Grossberg, S. (1980c). Biological competition: Decision rules, pattern formation, and oscillations. *Proceedings of the National Academy of Sciences*, **77**, 2338–2342.

Grossberg, S. (Ed.) (1981). Adaptive resonance in development, perception, and cognition. In S. Grossberg (Ed.), **Mathematical psychology and psychophysiology**. Providence, RI: American Mathematical Society.

Grossberg, S. (1982a). **Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control**. Boston, MA: Reidel Press.

Grossberg, S. (1982b). Associative and competitive principles of learning and development: The temporal unfolding and stability of STM and LTM patterns. In S.-I. Amari and M. Arbib (Eds.), **Competition and cooperation in neural networks**. New York: Springer-Verlag.

Grossberg, S. (1982c). A psychophysiological theory of reinforcement, drive, motivation, and attention. *Journal of Theoretical Neurobiology*, **1**, 286-369.

Grossberg, S. (1983). The quantized geometry of visual space: The coherent computation of depth, form, and lightness. *Behavioral and Brain Sciences*, **6**, 625–657.

Grossberg, S. (1984). Some psychophysiological and pharmacological correlates of a developmental, cognitive, and motivational theory. In J. Cohen, R. Karrer, and P. Tueting (Eds.), **Brain and information: Event related potentials, 425**, 58–151, Annals of the New York Academy of Sciences. Reprinted in S. Grossberg (Ed.), **The adaptive brain, Volume I**, 1987, Amsterdam: Elsevier/North-Holland.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E.C. Schwab and H.C. Nusbaum (Eds.), **Pattern recognition by humans and machines, Volume 1: Speech perception**, pp. 187–294, New York, NY: Academic Press. Reprinted in S. Grossberg (Ed.), **The adaptive brain, Volume II**, 1987, Amsterdam: Elsevier/North-Holland.

Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, **1**, 17–61.

Grossberg, S. and Kuperstein, M. (1986). **Neural dynamics of adaptive sensory-motor control**. Amsterdam: Elsevier/North-Holland; expanded edition, 1989, Elmsford, NY: Pergamon Press.

Grossberg, S. and Merrill, J.W.L. (1992). A neural network model of adaptively timed reinforcement learning and hippocampal dynamics. *Cognitive Brain Research*, **1**, 3–38.

Grossberg, S. and Mingolla, E. (1985a). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychological Review*, **92**, 173–211.

Grossberg, S. and Mingolla, E. (1985b). Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. *Perception and Psychophysics*, 1985, **38**, 141–171.

Grossberg, S. and Pepe, J. (1970). Schizophrenia: Possible dependence of associational span, bowing, and primacy versus recency on spiking threshold. *Behavioral Science*, **15**, 359–362.

Grossberg, S. and Pepe, J. (1971). Spiking threshold and overarousal effects in serial learning. *Journal of Statistical Physics*, **3**, 95–125.

Grossberg, S. and Somers, D. (1991). Synchronized oscillations during cooperative feature linking in a cortical model of visual perception. *Neural Networks*, **4**, 453–466.

Grossberg, S. and Somers, D. (1992). Synchronized oscillations for binding spatially distributed feature codes into coherent spatial patterns. In G.A. Carpenter and S. Grossberg, (Eds.), **Neural networks for vision and image processing**. Cambridge, MA: MIT Press, 385–406.

Grossberg, S. and Stone, G.O. (1986). Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review*, **93**, 46–74.

Grossberg, S. and Todorović, D. (1988). Neural dynamics of 1-D and 2-D brightness perception: A unified model of classical and recent phenomena. *Perception and Psychophysics*, **43**, 241–277.

Hebb, D.O. (1949). **The organization of behavior**. New York, NY: Wiley Press.

Hecht-Nielsen, R. (1987). Counterpropagation networks. *Applied Optics*, **26**, 4979–4984.

Hirsch, M.W. (1982). Systems of differential equations which are competitive or cooperative, I: Limit sets. *SIAM Journal of Mathematical Analysis*, **13**, 167–179.

Hirsch, M.W. (1985). Systems of differential equations which are competitive or cooperative, II: Convergence almost everywhere. *SIAM Journal of Mathematical Analysis*, **16**, 423–439.

Hirsch, M.W. (1989). Convergent activation dynamics in continuous time networks. *Neural Networks*, **2**, 331–350.

Hodgkin, A.L. (1964). **The conduction of the nervous system.** Liverpool, UK: Liverpool University.

Holland, J.H. (1980). Adaptive algorithms for discovering and using general patterns in growing knowledge bases. *International Journal of Policy Analysis and Information Systems*, **4**, 217–240.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, **79**, 2554–2558.

Hopfield, J.J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, **81**, 3058–3092.

Hubel, D.H. and Wiesel, T.N. (1977). Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London (B)*, **198**, 1–59.

Hunt, R.K. and Jacobson, M. (1974). Specification of positional information in retinal ganglion cells of *Xenopus laevis*: Intraocular control of the time of specification. *Proceedings of the National Academy of Sciences*, **71**, 3616–3620.

Iverson, G.J. and Pavel, M. (1981). Invariant properties of masking phenomena in psychoacoustics and their theoretical consequences. In S. Grossberg (Ed.), **Mathematical psychology and psychophysiology**. Providence, RI: American Mathematical Society, pp. 17–24.

Jung, J. (1968). **Verbal learning.** New York: Holt, Rinehart, and Winston.

Kandel, E.R. and O'Dell, T.J. (1992). Are adult learning mechanisms also used for development? *Science*, **258**, 243–245.

Kandel, E.R. and Schwartz, J.H. (1981). **Principles of neural science.** New York, NY: Elsevier/North-Holland.

Katz, B. (1966). **Nerve, muscle, and synapse.** New York, NY: McGraw-Hill.

Khinchin, A.I. (1967). **Mathematical foundations of information theory.** New York, NY: Dover Press.

Klatsky, R.L. (1980), **Human memory: Structures and processes.** San Francisco, CA: W.H. Freeman.

Kohonen, T. (1984). **Self-organization and associative memory**, New York, NY: Springer-Verlag.

Kosko, B. (1986). Fuzzy entropy and conditioning. *Information Sciences*, **40**, 165-174.

Levine, D. and Grossberg, S. (1976). On visual illusions in neural networks: Line neutralization, tilt aftereffect, and angle expansion. *Journal of Theoretical Biology*, **61**, 477–504.

Levy, W.B. (1985). Associative changes at the synapse: LTP in the hippocampus. In W.B. Levy, J. Anderson and S. Lehmkuhle, (Eds.), **Synaptic modification, neuron selectivity, and nervous system organization.** Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 5–33.

Levy, W.B., Brassel, S.E., and Moore, S.D. (1983). Partial quantification of the associative synaptic learning rule of the dentate gyrus. *Neuroscience*, **8**, 799–808.

Levy, W.B. and Desmond, N.L. (1985). The rules of elemental synaptic plasticity. In W.B. Levy, J. Anderson and S. Lehmkuhle, (Eds.), **Synaptic modification, neuron**

selectivity, and nervous system organization. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 105-121.

Linsker, R. (1986). From basic network principles to neural architecture. *Proceedings of the National Academy of Science*, **83**, 7508-7512, 8390–8394, 8779–8783.

Maher, B.A. (1977). **Contributions to the psychopathology of schizophrenia.** New York, NY: Academic Press.

Malsburg, C. von der (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, **14**, 85–100.

Malsburg, C. von der and Willshaw, D.J. (1981). Differential equations for the development of topological nerve fibre projections. In S. Grossberg (Ed.), **Mathematical psychology and psychophysiology.** Providence, RI: American Mathematical Society, pp. 39–48.

May, R.M. and Leonard, W.J. (1975). Nonlinear aspects of competition between three species. *SIAM Journal on Applied Mathematics*, **29**, 243–253.

McCulloch, W.S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of the Mathematical Biophysics*, **5**, 115–133.

McGeogh, J.A. and Irion, A.L. (1952). **The psychology of human learning, Second edition.** New York: Longmans and Green.

Miller, G.A. (1956). The magic number seven plus or minus two. *Psychological Review*, **63**, 81.

Moore, B. (1989). ART 1 and pattern clustering. In D. Touretzky, G. Hinton, and T. Sejnowski (Eds.), **Proceedings of the 1988 connectionist models summer school.** San Mateo, CA: Morgan Kaufmann, pp. 174-185.

Murdock, B.B. (1974). **Human memory: Theory and data.** Potomac, MD: Erlbaum Press.

Nabet, B. and Pinter, R.B. (1991). **Sensory neural networks: Lateral inhibition.** Boca Raton, FL: CRC Press.

Norman, D.A. (1969). **Memory and attention: An introduction to human information processing.** New York, NY: Wiley and Sons.

Osgood, C.E. (1953). **Method and theory in experimental psychology.** New York, NY: Oxford Press.

Plonsey, R. and Fleming, D.G. (1969). **Bioelectric phenomena.** New York, NY: McGraw-Hill.

Rauschecker, J.P. and Singer, W. (1979). Changes in the circuitry of the kitten's visual cortex are gated by postsynaptic activity. *Nature*, 280, 58–60.

Repp, B.H. (1991). Perceptual restoration of a "missing" speech sound: Auditory induction or illusion? **Haskins Laboratories Status Report on Speech Research,** SR-107/108, 147–170.

Rumelhart, D.E. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, **9**, 75–112.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, **89**, 63–77.

Salzberg, S.L. (1990). **Learning with nested generalized exemplars.** Boston, MA: Kluwer Academic Publishers.

Samuel, A.G. (1981a). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, **110**, 474–494.

Samuel, A.G. (1981b). The rule of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 1124–1131.

Schvaneveldt, R.W. and MacDonald, J.E. (1981). Semantic context and the encoding of words: Evidence for two modes of stimulus analysis. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 673–687.

Singer, W., Neuronal activity as a shaping factor in the self-organization of neuron assemblies. In E. Basar, H. Flohr, H. Haken, and A.J. Mandell (Eds.) (1983). **Synergetics of the brain.** New York, NY: Springer-Verlag, pp. 89–101.

Smith, E.E. (1990). In D.O. Osherson and E.E. Smith (Eds.), **An invitation to cognitive science.** Cambridge, MA: MIT Press.

Somers, D. and Kopell, N. (1993). Rapid synchronization through fast threshold modulation. *Biological Cybernetics*, in press.

Underwood, B.J. (1966). **Experimental psychology, Second edition.** New York: Appleton-Century-Crofts.

Warren, R.M. (1984). Perceptual restoration of obliterated sounds. *Psychological Bulletin*, **96**, 371–383.

Warren, R.M. and Sherman, G.L. (1974). Phonemic restorations based on subsequent context. *Perception and Psychophysics*, **16**, 150–156.

Werblin, F.S. (1971). Adaptation in a vertebrate retina: Intracellular recordings in *Necturus*. *Journal of Neurophysiology*, **34**, 228–241.

Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. Ph.D. Thesis, Cambridge, MA: Harvard University.

Willshaw, D.J. and Malsburg, C. von der (1976). How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London (B)*, **194**, 431–445.

Young, R.K. (1968). Serial learning. In T.R. Dixon and D.L. Horton (Eds.), **Verbal behavior and general behavior theory.** Englewood Cliffs, NJ: Prentice-Hall.

Zadeh, L. (1965). Fuzzy sets. *Information Control*, **8**, 338–353.